Classification of Alzheimer's Disease via Vision Transformer

Classification of Alzheimer's Disease via Vision Transformer

Yanjun Lyu
Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, Texas, USA
yxl9168@mavs.uta.edu

Xiaowei Yu
Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX, USA
xxy1302@mavs.uta.edu

Dajiang Zhu
Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX, USA
dajiang.zhu@uta.edu

Lu Zhang
Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX, USA
lu.zhang2@mavs.uta.edu

and for the Alzheimer's Disease Neuroimaging Initiative*
Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu).

ABSTRACT

Deep models are powerful in capturing the complex and non-linear relationship buried in brain imaging data. However, the huge number of parameters in deep models can easily overfit given limited imaging data samples. In this work, we proposed a cross-domain transfer learning method to solve the insufficient data problem in brain imaging domain by leveraging the knowledge learned in natural image domain. Specifically, we employed ViT as the backbone and firstly pretrained it using ImageNet-21K dataset and then transferred to the brain imaging dataset. A slice-wise convolution embedding method was developed to improve the standard patch operation in vanilla ViT. Our method was evaluated based on AD/CN classification task. We also conducted extensive experiments to compare the transfer performance with different transfer strategies, models, and sample size. The results suggest that the proposed method can effectively transfer the knowledge learned in natural image domain to brain imaging area and may provide a promising way to take advantages of the pretrained model in data-intensive applications. Moreover, the proposed cross-domain transfer learning method can obtain comparable classification performance compared to most recent studies.

CCS CONCEPTS

• **Computing methodologies**; • **Machine learning**; • **Machine learning approaches**; • **Neural networks**;

KEYWORDS

Transfer learning, Cross-domain, AD classification, Brain image

ACM Reference Format:
Yanjun Lyu, Xiaowei Yu, Dajiang Zhu, Lu Zhang, and and for the Alzheimer's Disease Neuroimaging Initiative*. 2022. Classification of Alzheimer's Disease via Vision Transformer: Classification of Alzheimer's Disease via Vision Transformer. In *The15th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '22), June 29–July 01, 2022, Corfu, Greece.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3529190.3534754

1 INTRODUCTION

In previous studies, deep learning has demonstrated breakthroughs of performance in neuroimaging analysis [1-7]. Deep models can be especially useful in capturing the complex and non-linear relationship buried in brain imaging data. Recent development of transformer-based deep models has revolutionized the field of deep learning. For example, vision transformer (ViT) [12] has shown its superiority in many studies. Different from convolution neural networks (CNNs) that aggregate features gradually from local to global by stacking more convolutional layers, ViT takes advantages of the multi-headed self-attention mechanism to capture the long-range dependencies which allows the model to attend over all elements in the input sequence and thus achieves better performance. As shown in [8], a ViT model with 22M learnable parameters can achieve better performance than the ResNet-101 model which has more than 30 bottleneck convolutional blocks in ImageNet classification task. This characteristic of ViT is ideal for brain imaging analysis, as brain is a highly complex network where brain regions far away from each other may have strong relationships. The self-attention mechanism of ViT can effectively capture the dependencies among remoted brain regions.

The continuously growing huge architectures have enabled the ViT to achieve great success. However, the huge number of parameters begin to demand hundreds of millions of labeled data which are often publicly inaccessible in brain imaging domain. A

PETRA '22, June 29–July 01, 2022, Corfu, Greece

ACM ISBN 978-1-4503-9631-8/22/06…$15.00
https://doi.org/10.1145/3529190.3534754

promising method to tackle this problem is the transfer learning [9]. Inspired by human beings' capabilities to intelligently apply knowledge learned previously to solve new problems faster and better, transfer learning aims to leverage knowledge learned from a related domain (source domain) to improve the performance in a target domain. Transfer learning has the potential to alleviate the problem of insufficient data in brain imaging domain by leveraging the knowledge learned in another data-intensive domain, such as natural images domain. Generally, there are two main factors heavily influencing the effectiveness of transfer learning. One is the relevance between the source and the target domains and the other is the model's capacity of characterizing the transferable part of the knowledge across domains. As brain imaging and natural images are both image data, the two types of image data share a lot of basic image features such as edge and shape, which provides a foundation for transfer learning. By applying a powerful learner, such as ViT, the knowledge learned in natural images can be adapted to brain imaging data well.

In this work, we proposed a cross-domain transfer learning method to solve the insufficient data problem in brain imaging domain. Specifically, we employed ViT as the backbone. The ViT model was firstly pretrained using ImageNet-21K dataset and then transferred to the brain imaging dataset. Inspired by recent work that introduces convolution layers into ViT [10, 11], a slice-wise convolution embedding method was employed in this work. We evaluated the proposed method based on AD/CN classification task using different transfer strategies. Extensive experiments have been conducted to compare the transfer performance of different models and to evaluate the influence of sample size on the model performance. The results suggest that the proposed method can effectively transfer the knowledge learned in natural image domain to brain imaging area, which may provide a promising way for the application areas with very limited data samples to take advantages of the pretrained model in data-intensive applications. Moreover, the proposed cross-domain transfer learning method can obtain comparable classification performance compared to most recent studies.

2 METHOD

2.1 Overview

In this work, we conducted research on cross-domain transfer learning. Taking Vision-Transformer (ViT) as backbone [12], we first trained ViT using natural images to take full advantages of the large-scale data in data-intensive computer vision domain, and then the pretrained model was transferred to the brain imaging domain where publicly accessible samples are very limited. In this section, the details of the proposed method will be introduced and organized as follows: we first introduced the data collection and pre-processing pipeline of brain imaging data in Section 2.2; Then, the architecture of ViT will be introduced in Section 2.3; Finally, in Section 2.4 we illustrated the details of fitting brain imaging to the pretrained ViT model by convolutional patch-wise embedding, which is the key in the proposed cross-domain transfer learning. The model was trained by AD/CV classification task and the details will be discussed in Section 2.4.

2.2 Data collection and preprocessing

In this work, the ViT was first pretrained using ImageNet-21K dataset which includes 1.2M natural images that belong to 1,000 mutually exclusive classes. Then the model was transferred to the brain imaging dataset, where structure MRI (T1 weight) data of 505 subjects (284 CN/221 AD) were collected from Alzheimer's Disease Neuroimaging Initiative (ADNI) [13]. A quality control step was conducted to match the meta information between the two clinical groups, e.g., age and gender. The brain imaging data with low quality have been excluded. After the quality control, we obtained 375 subjects (265 CN/110 AD) in total.

2.2.1 Image data processing. The imaging parameters for T1-weight MRI are: TR=2300.0ms, TE=3.0ms, image matrix= 240×256×208, with resolution of 1.0×1.0×1.0mm^3. We applied the same standard pre-processing procedures as in [4] for T1 imaging data. In brief, pre-processing steps include brain skull removal, linear registration via FLIRT to warp T1 imaging with the MNI 152 template in standard MNI space. After that, a cropping step was applied to remove the background and as a result, the images were resized into 140×150×100. The resized images were further down sampled into 70×75×50 and normalized by Z-normalization. Finally, the whole brain was organized by slice-wise manner from coronal direction, that is: for each subject, there are 75 slices with dimension of 70×50. Each slice was assigned the same label as the corresponding subject and used as a data sample to train the model.

2.3 Vision Transformer

The architecture of Vit used in this work is depicted in Figure 1(a). The key component of ViT model is the transformer encoder [14]. As shown in Fig. 1, transformer encoder is composed of a stack of multiple identical layers. Each layer has two sub-layers, one is the multi-head attention layer and the other is the multi-layer perceptron (MLP). The MLP contains two layers with a GELU non-linearity. A residual connection [15] is employed around each of the two sub-layers and followed by layer normalization (Lnorm) [16]. The input (Z_0) of the transformer is a sequence of N embedded image patches (T_p^i) and a special token T_{cls}. The state of the special token at the output (Z_l^0) of the transformer encoder can serve as the image representation y used for the classification task. The learnable position embeddings are added to each of the patch embedding as well as the special token in the input sequence. The pipeline of the Vit can be formulated by Eq. 1-Eq. 4:

$$\begin{aligned} Z_0 \quad = \quad & \left[T_{cls}\,; T_p^1\,; T_p^2; \ldots,\,; T_p^N\right] \\ & + T_{pos}\,,\quad T_{cls},\, T_p^i \in R^{1\times d},\quad T_{pos} \in R^{(N+1)\times d} \end{aligned} \tag{1}$$

$$Z_\ell^* = MSA\left(Lnorm\left(Z_{\ell-1}\right)\right) + \, Z_{\ell-1} \tag{2}$$

$$Z_\ell = MLP\left(Lnorm\left(Z_\ell^*\right)\right) + \, Z_\ell^* \tag{3}$$

$$y = Lnorm\left(Z_\ell^0\right) \tag{4}$$

where the T_p^i is the patch embedding, T_{cls} is the special classification token, T_{pos} is the position embedding. MSA denotes the multi-head attention layers. Z_l is the output logit of block l, and Z_l^0 is the state of classification token after block l . Lnorm denotes the layer-norm operation. In vanilla Vit, the patch embedding is

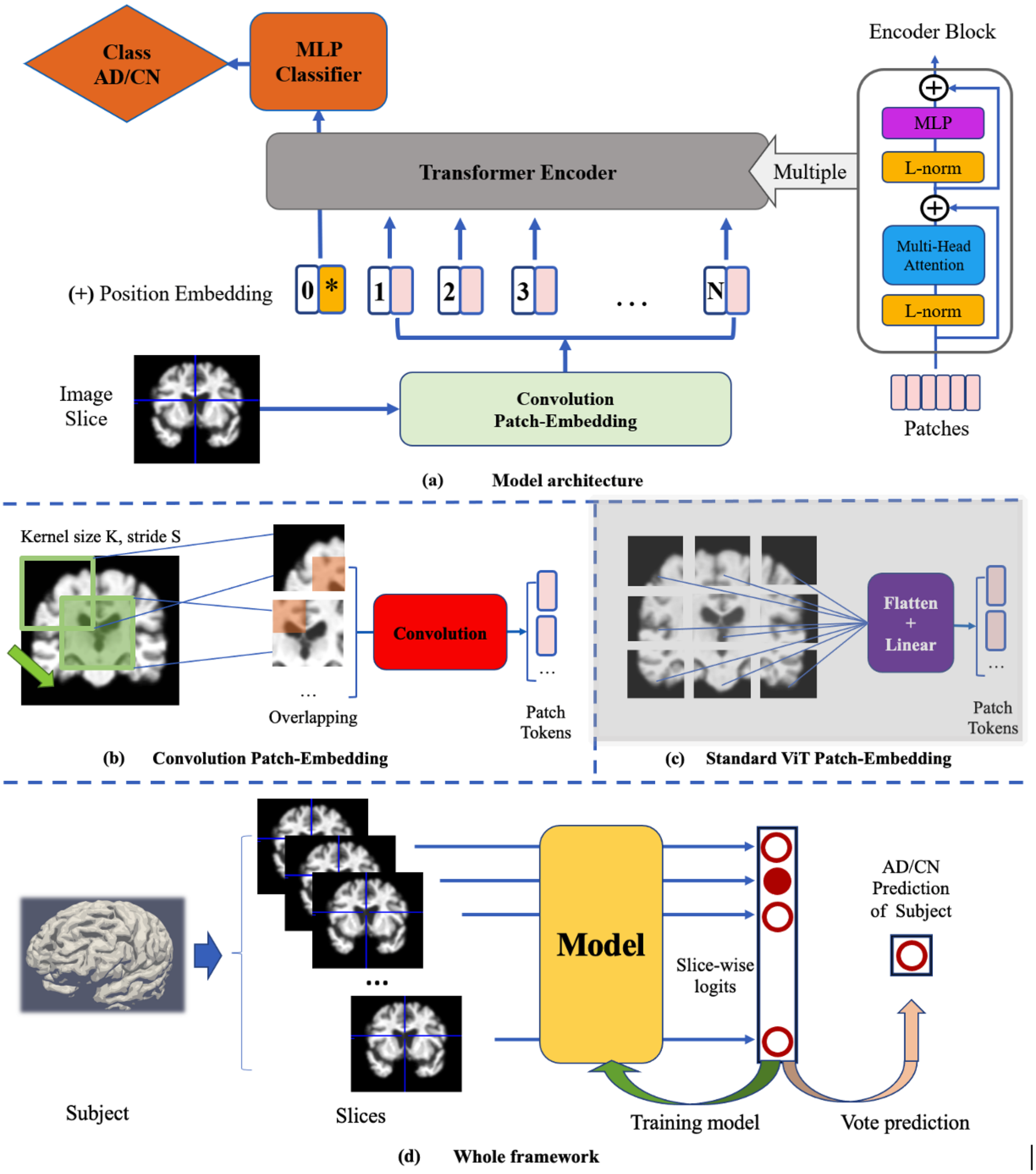

Figure 1: Illustration of the proposed cross-domain transfer learning based on ViT model. (a) The model architecture: each slice of brain image was treated as the input of a patch-wise embedding block by a convolution operation. The special CLS token (denoted as *) is used to make the prediction. The whole transformer encoder is composed of multiple Transformer blocks (right). (b) The details of convolution patch embedding. The embedding is implemented by a convolution layer kernel size K and stride size S. (c) Standard ViT patch-embedding (for comparison only, faded as not been applied in this work). (d) The framework of the training phase and prediction phase of the model in the classification task.

implemented by first flatting the image patches and then mapping them to d dimensions embedding space with a trainable linear projection (Figure 1(c)). In this work, to adapt the pretrained ViT model in brain imaging domain, we adopted a modified patch-wise embedding layer (Figure 1(b)), which will be elaborated in Section 2.4.

2.4 Patch-wise embedding and slice-wise operations

2.4.1 Brain image patch-wise embedding. To adapt the pretrained ViT model in brain imaging domain, we organized the brain imaging data into slices and conducted patch-wise embedding. Specifically, we first reshaped the 3D brain volume of one subject into H 2D slices (H is the coronally dimension). Each slice was considered as a new sample and assigned the same label as the corresponding subject. This process can be formulated as: $V \in R^{W\times H\times G\times C} \rightarrow [x_i \in R^{W\times G\times C}]_{i=1}^{H}$, where V is the original 3D volume and x is a single slice.

The encoder of ViT model requires the input as sequence of tokens [12]. A patch embedding layer is needed to transform any high-dimensional data into a sequence of embedding vectors. Specifically, in this work, a 2D slices $x \in R^{W\times G\times C}$ will be reshaped as $x_p \in R^{N\times(p^2\times C)}$, where C=1 is the input channel, (W, G) is the input resolution, P is the patch resolution and N is the number of patches. Inspired by previous study [10, 11], convolution operation with overlapping (Figure 1(b)) has replaced the rigid patch-wise operation in vanilla ViT to avoid directly using the non-overlapping image patches as input, Figure 1(c). We set the kernel size of the convolution layer to be K and set the stride to be S ($S < K$), to enable the overlapping between patches, and the output channel size of the convolution layer is corresponding to the embedding dimension of the transformer blocks. In the convolution operation, the whole slice x was divided into N patches, $N = (\frac{W+2E-K}{S} + 1) \times (\frac{G+2E-K}{S} + 1)$, where E is a flexible padding size to make N to be an integer. The learnable kernel will replace the linear layer in vanilla ViT to project the patches into token-vectors (Figure 1(b)-(c)). Same with the standard ViT, we also adopted 1D learnable position embeddings to record the positional information of patches.

2.4.2 Slice-wise training and prediction. The proposed model was applied to the AD/CN classification task. In the training process, each slice with the class label was considered as a data sample and the model generated a predicted label for each input slice. The predicted label logit was then compared with the ground truth to optimize the model via cross-entropy loss. In the prediction process, the model will output a predict logit for each input slice x, and the final prediction of each subject come from the vote result of the set of H slices.

3 EXPERIMENTS AND RESULTS

To evaluate the effectiveness of cross-domain pretrained ViT, we conducted extensive experiments based on AD/CN classification task. In section 3.2.1, we compared the transfer performance of different models including three kinds of ViT with different model size and two different architectures of ResNet family. In section 3.2.2, we evaluated the influence of the size of brain imaging dataset on the transfer performance. In section 3.2.3, we compared the classification performance of the proposed method with three most recent studies about AD/CN classification problem. The model setting will be introduced in section 3.1.

3.1 Experimental setting

3.1.1 Different models and training strategies. We compared three homogeneous ViT which have different model size (tiny, small, base), and two ResNet architectures including ResNet18 and ResNet34 (Section 3.2.1). These models have been reported in previous studies to obtain good performance in capturing image features [17]. For each of these models, we compared three different types of training strategies:

Fine-tune only method (FT): during the training process, the parameters of the transformer encoder will remain unmodified and only the patch embedding layers and classifier will be updated according to the AD/CN classification loss.

Train-from-scratch method (Scratch): the parameters of the whole model will be randomly initialized and updated during the whole training process.

Training from checkpoint method (Check-Point): during the first 50 epochs of the training process, the parameters of the transformer encoder will be frozen and the patch embedding layers and classifier will be updated. After the first 50 epochs, the frozen layers will be freed and the weights will be updated with a small learning rate.

Considering the two classes (AD/CN) is imbalanced in our dataset, we applied a weighted sampler in training process to make every training batch balanced.

3.1.2 Different sample size. To evaluate the influence of sample size on the transfer performance, we conducted ablation study in Section 3.2.2 using ViT small model architecture which obtained the best performance compared to the other four (ViT tiny, ViT base, ResNet18, ResNet34). Specifically, we randomly selected three sub-dataset of different sample size from the whole dataset, including dataset-1 with AD/CN=40/40, dataset-2 with AD/CN=80/80, and dataset-3 with AD/CN=100/100. For each sub-dataset, we compared the three training strategies (Ft, Scratch, Check-Point in section 3.1.1) with a splitting of training/validation/testing equal to 60%/20%/20%.

3.1.3 Hyper-parameters. In our experiments, the convolutional layers of Resnet18 and Resnet34 adopted the kernel size of 7 and stride of 3. The training process of different models is implemented under the same hyper-parameter setting as following: initial learning rate 0.001, total epoch 120, dropout rate of 0.3 for linear layers. The optimizer is Adam and a 'StepLR' scheduler (gamma is 0.5, step is 10 epochs) was applied to decrease the learning rate during the training process. We reported the performance of different model settings on testing dataset in the section 3.2 and the classification accuracy is chosen as the metric of model evaluating.

3.2 Classification performance

3.2.1 Classification performance of different models and training strategies. In this section we reported and analyzed the classification performance of the five different model settings introduced in

Table 1: Classification Performance of Different Models and Training Strategies (FT: fine-tuning, Check-Point: continue train the model from pretrained checkpoint, Scratch: randomly initial the model and train from scratch)

Model	Training	ACC (Avg.)	Recall	Precision	F1	Model Size	Embedding dim
ViT tiny	FT	95.3%	94.4%	90.0%	0.932	5.7M	192
ViT tiny	Check-Point	94.8%	92.4%	90.1%	0.911	5.7M	192
ViT tiny	Scratch	91.6%	84.9%	86.7%	0.857	5.7M	192
ViT small	FT	96.8%	93.3%	96.7%	0.949	22M	384
ViT small	Check-Point	96.2%	91.0%	97.1%	0.939	22M	384
ViT small	Scratch	90.6%	95.4%	83.3%	0.889	22M	384
ViT base	FT	93.7%	93.3%	91.0%	0.921	86M	768
ViT base	Check-Point	95.0%	95.3%	95.3%	0.953	86M	768
ViT base	Scratch	88.9%	89.8%	84.8%	0.872	86M	768
ResNet18	FT	84.8%	100.0%	66.6%	0.799	11M	-
ResNet18	Check-Point	88.5%	100.0%	75.9%	0.863	11M	-
ResNet18	Scratch	71.7%	88.8%	52.3%	0.658	11M	-
ResNet34	FT	75.9%	100.0%	55.3%	0.712	21M	-
ResNet34	Check-Point	75.9%	100.0%	55.3%	0.712	21M	-
ResNet34	Scratch	68.1%	88.9%	50.1%	0.641	21M	-

Section 3.1.1. For each model setting, we adopted the three training strategies and showed the results in Table 1. From Table 1 we can see that within each model the two kinds of transfer-based strategies (FT and Check-Point) obtain better results than training from the scratch. This result suggests that the proposed cross-domain transfer learning method can effectively adapt the learned knowledge from natural images to brain imaging data. The study of this work may provide a promising way for the application areas with very limited data samples to take advantages of the pretrained model in data-intensive applications. In addition, all the three ViT models with different model sizes outperform the two CNN based ResNet architectures, which indicates the superior performance of the ViT models. It is also noteworthy that the 100% recall values together with much smaller precision values appear in the results obtained by ResNet models. This means the model assigned the same label for samples from different classes and failed in the classification task. This might because CNN based deep models cannot directly capture the long-range dependencies of image features and hence the special features of the natural images cannot adapt quickly to the brain imaging data. Moreover, compared to the ViT tiny and ViT base, ViT small obtained slightly higher accuracy when trained by FT strategy. It might because that compared to ViT tiny, ViT small with a larger model size can capture more complicated relationship buried in the brain imaging data and hence can obtain better classification performance. But a large model such as ViT base may suffer from overfitting problem and harm the performance.

3.2.2 Evaluation of the influence of sample size to the classification performance. Transfer learning can alleviate the problem of insufficient training samples. However, the sample size still has influence on the transfer performance. In this section we evaluated the influence of the sample size using ViT small model, which obtained the best performance in our previous experiments summarized in Table 1. As shown in Table 2, when the sample size decreases, the accuracy of two transfer-related methods decreased slightly, while the accuracy of the training from scratch method decreased dramatically. This result suggests that compared to training from scratch, transfer learning is less sensitive to sample size.

3.2.3 Comparison with related works. In this section, we compared the proposed cross-domain transfer learning method with previous studies. For fairly comparison, we summarize the overall classification performance of recent studies on AD/CN classification task using T1 weighted MRI data from ADNI dataset. The results have been reported in Table 1. From the results we can see that, the proposed cross-domain transfer learning method can obtain comparable classification performance on AD/CN classification task using a smaller training dataset.

4 CONCLUSION

In this work we proposed a cross-domain transfer learning method to solve the insufficient data problem in brain imaging domain. We employed ViT as the backbone and firstly pretrained it using ImageNet-21K dataset, then transferred to the brain imaging dataset. We evaluated the proposed method based on AD/CN classification task with extensive comparisons to different transfer strategies, models, and sample size. The results suggest that our method can effectively transfer the knowledge learned in natural image domain to brain imaging area. Moreover, the proposed cross-domain transfer learning method can obtain comparable classification performance compared to most recent studies using a smaller training dataset.

REFERENCES

[1] Lu Zhang, Li Wang, and Dajiang Zhu., 2020, Recovering Brain Structural Connectivity from Functional Connectivity via Multi-GCN Based Generative Adversarial Network. In Medical Image Computing and Computer Assisted Intervention – MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VII. Springer-Verlag, Berlin, Heidelberg, 53–61. https://doi.org/10.1007/978-3-030-59728-3_6

[2] Zhang, L., Wang, L., & Zhu, D., 2020, Jointly Analyzing Alzheimer's Disease Related Structure-Function Using Deep Cross-Model Attention Network. 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), 563-567..

Table 2: Classification Performance of ViT Small Model with Different Sample Size (FT: fine-tuning, Check-Point: continue train the model from pretrained checkpoint, Scratch: randomly initial the model and train from scratch)

CN/AD	Train/Val/Test	Training	ACC (Avg.)	Recall	Precision	F1
100/100	120/40/40	FT	96.8%	97.2%	97.2%	0.972
100/100	120/40/40	Check-Point	92.4%	94.4%	91.6%	0.929
100/100	120/40/40	Scratch	84.3%	93.7%	80.5%	0.866
80/80	96/32/32	FT	93.3%	100.0%	88.9%	0.941
80/80	96/32/32	Check-Point	91.8%	91.6%	95.0%	0.932
80/80	96/32/32	Scratch	73.6%	60.0%	91.6%	0.725
60/60	72/24/24	FT	90.9%	100.0%	85.7%	0.922
60/60	72/24/24	Check-Point	92.7%	100.0%	88.9%	0.941
60/60	72/24/24	Scratch	36.4%	16.0%	75.0%	0.223

Table 3: Comparison of Classification Performance with Recent Studies

Work	Model	Modality	Sample Size	ACC (Avg.)	F1
Liu et al. (2019) [18]	DM2L	MRI image	181 AD, 226 CN	93.7%	-
Basaia et al. (2019) [19]	Analysis & CNNs	MRI image	294 AD, 352 CN	99.2%	-
C. Lian et al. (2020) [20]	H-FCN	MRI image	199 AD, 229 CN	90.3%	-
Ours	Vit fine-tune	MRI image	110 AD, 265 CN	96.8%	0.949

[3] Zhang, L., Zaman, A., Wang, L., Yan, J. and Zhu, D., 2019, October. A Cascaded Multi-Modality Analysis in Mild Cognitive Impairment. In International Workshop on Machine Learning in Medical Imaging (pp. 557-565). Springer, Cham.

[4] Zhang, L., Wang, L., Gao, J., Risacher, S.L., Yan, J., Li, G., Liu, T., Zhu, D. and Alzheimer's Disease Neuroimaging Initiative, 2021. Deep fusion of brain structure-function in mild cognitive impairment. Medical image analysis, 72, p.102082.

[5] Wang, L., Zhang, L. and Zhu, D., 2020, April. Learning Latent Structure Over Deep Fusion Model of Mild Cognitive Impairment. In 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI) (pp. 1039-1043). IEEE.

[6] Zaman, A., Zhang, L., Yan, J. and Zhu, D., 2019, October. Multi-modal Image Prediction via Spatial Hybrid U-Net. In International Workshop on Multiscale Multimodal Medical Imaging (pp. 1-9). Springer, Cham.Prokop, Emily. 2018. The Story Behind. Mango Publishing Group. Florida, USA.

[7] Wang, L., Zhang, L. and Zhu, D., 2019, April. Accessing Latent Connectome of Mild Cognitive Impairment via Discriminant Structure Learning. In 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019) (pp. 164-168). IEEE

[8] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Herve J ´ egou. Training ´ data-efficient image transformers & distillation through attention. arXiv preprint arXiv:2012.12877, 2020.

[9] Pan, S.J. and Yang, Q., 2009. A survey on transfer learning. IEEE Transactions on knowledge and data engineering, 22(10), pp.1345-1359.

[10] Wu, Haiping and Xiao, Bin and Codella, Noel and Liu, Mengchen and Dai, Xiyang and Yuan, Lu and Zhang, Lei. 2021. CvT: Introducing Convolutions to Vision Transformers, Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 22-31.

[11] Tete Xiao and Piotr Dollar and Mannat Singh and Eric Mintun and Trevor Darrell and Ross Girshick, 2021, Early Convolutions Help Transformers See Better, Advances in Neural Information Processing Systems.

[12] Alexey Dosovitskiy and Lucas Beyer and Alexander Kolesnikov and Dirk Weissenborn and Xiaohua Zhai and Thomas Unterthiner and Mostafa Dehghani and Matthias Minderer and Georg Heigold and Sylvain Gelly and Jakob Uszkoreit and Neil Houlsby, 2020, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, CoRR, abs/2010.11929

[13] ADNI | Alzheimer's Disease Neuroimaging Initiative, http://adni.loni.usc.edu

[14] Vaswani, Ashish and Shazeer, Noam and Parmar, Niki and Uszkoreit, Jakob and Jones, Llion and Gomez, Aidan N and Kaiser, L ukasz and Polosukhin, Illia, 2017, Attention is All you Need, Advances in Neural Information Processing Systems.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 770–778, 2016.

[16] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. arXiv preprint arXiv:1607.06450, 2016

[17] He, Kaiming and Zhang, Xiangyu and Ren, Shaoqing and Sun, Jian, 2016, Deep Residual Learning for Image Recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770-778

[18] Liu M, Zhang J, Adeli E, Shen D. Joint Classification and Regression via Deep Multi-Task Multi-Channel Learning for Alzheimer's Disease Diagnosis. IEEE Trans Biomed Eng. 2019 May;66(5):1195-1206. doi: 10.1109/TBME.2018.2869989. Epub 2018 Sep 12. PMID: 30222548; PMCID: PMC6764421.

[19] Basaia S, Agosta F, Wagner L, Canu E, Magnani G, Santangelo R, Filippi M; Alzheimer's Disease Neuroimaging Initiative. Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks. Neuroimage Clin. 2019;21:101645. doi: 10.1016/j.nicl.2018.101645. Epub 2018 Dec 18. PMID: 30584016; PMCID: PMC6413333.

[20] C. Lian, M. Liu, J. Zhang and D. Shen, 2020, Hierarchical Fully Convolutional Network for Joint Atrophy Localization and Alzheimer's Disease Diagnosis Using Structural MRI, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 4, pp. 880-893, doi: 10.1109/TPAMI.2018.2889096.