Training ViT with Limited Data for Alzheimer's Disease Classification: an Empirical Study

Kassymzhomart Kunanbayev, Vyacheslav Shen, and Dae-Shik Kim

KAIST, 291 Daehak-ro, Yuseong-gu, Daejeon, South Korea
{kkassymzhomart, shen9910, daeshik}@kaist.ac.kr

Abstract. In this paper, we conduct an extensive exploration of a Vision Transformer (ViT) in brain medical imaging in a low-data regime. The recent and ongoing success of Vision Transformers in computer vision has motivated its development in medical imaging, but trumping it with inductive bias in a brain imaging domain imposes a real challenge since collecting and accessing large amounts of brain medical data is a labor-intensive process. Motivated by the need to bridge this data gap, we embarked on an investigation into alternative training strategies ranging from self-supervised pre-training to knowledge distillation to determine the feasibility of producing a practical plain ViT model. To this end, we conducted an intensive set of experiments using a small amount of labeled 3D brain MRI data for the task of Alzheimer's disease classification. As a result, our experiments yield an optimal training recipe, thus paving the way for Vision Transformer-based models for other low-data medical imaging applications. To bolster further development, we release our assortment of pre-trained models for a variety of MRI-related applications: https://github.com/qasymjomart/ViT_recipe_for_AD

Keywords: Vision Transformer · Alzheimer's Disease · Low-data regime

1 Introduction

The decade-long triumph of convolutional neural network-based (CNN) models in computer vision has been overtaken by Vision Transformer (ViT) [7] models. Despite having vision-specific inductive biases such as locality and spatial invariance [7], [1], CNN models fall short in accommodating long-range global dependencies – the feature that transformer-based models tend to naturally possess [25]. Long-range global dependencies are principally important in analyzing medical images, including brain MRI, the 3D high-dimensional nature of which requires precise focus on important discriminative locations. Furthermore, it is observed that 3D CNN models demand greater computational resources compared to the ViT models. Yet, to achieve remarkable performance, Vision Transformers remain data-hungry, i.e. they tend to require large amounts of data to secure inductive bias [7]. A substantial number of works have been focused on training Vision Transformers in a low-data regime [16], [18], [23], but only a handful of works focus on brain imaging applications [17].

Accordingly, the limited availability of structural brain MRI datasets, especially related to specific diseases like Alzheimer's disease (AD), may hamstring the development of accurate diagnostic tools. Although many prior works utilized CNN models with remarkable results, CNN-based approaches often downsample the input features, leading to the potential loss of information [21]. Other works that adopt Vision Transformers either combine with CNN feature extractors in a hybrid fashion or use large datasets for training [17], [14].

In this paper, we conducted empirical experiments with a plain ViT model to process a limited amount of 3D brain MRI data. The main reason for focusing on the plain version is to examine its potential without CNN-based feature extractors. By examining various training strategies and their effectiveness, our empirical experiments illustrate that a holistic orchestration of certain training strategies can boost the performance of a ViT model when trained with a few labeled samples. To summarize, our contributions can be outlined as follows:

- We pre-train the plain ViT model via self-supervised learning with the separate brain MRI data from different clinical cohorts and transfer it to fine-tune on AD classification task with few labeled data
- We explore the impact of pre-training with different pre-training data sizes and several non-homogeneous pre-training datasets
- In fine-tuning, we further examine the performance of the model in more extreme low-data scenarios
- We also investigate the application of different training methods such as 3D data augmentations, regularization, and knowledge distillation
- To bolster further research in the brain MRI community, we make available our assortment of pre-trained ViT models, trained with various combinations of non-homogeneous MRI datasets.

2 Method

2.1 Training strategies

Model architecture Encoder model architecture is important for processing brain MRI data and extracting global feature vector representations for further downstream applications, like classification. In this work, we experiment with the widely used configuration of the Vision Transformer −ViT-B [7], which has balanced computational efficiency and model complexity with 12 layers, 12 heads, a width of 768, and 86 million parameters. Since our input is 3D dimensional, we design our model for 3D input by replacing a linear 2D patch embedding with a 3D equivalent (see Fig. 1) with the same patch size of 16.

Pre-training has become a de facto one of the key ingredients in developing successful deep learning models [22], [9] and providing Vision Transformers with inductive bias [7]. Furthermore, given the prevalence of a low-data regime in medical imaging, particularly when related to diseases, we conjecture the importance of self-supervised pre-training using available medical data. To this end,

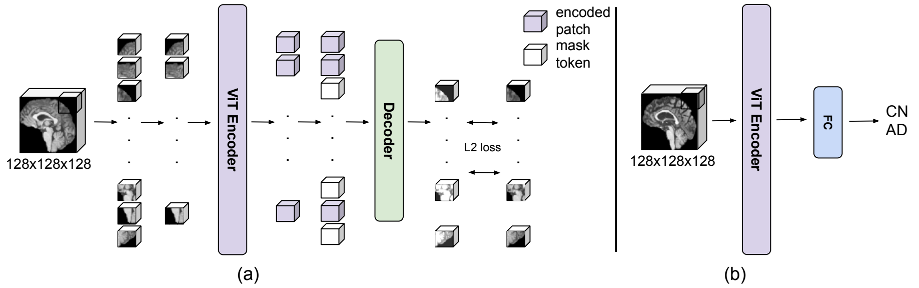

Fig. 1. Masked Autoencoders (MAE) pre-training (a) and fine-tuning (b) for 3D brain MRI data. For fine-tuning, pre-trained weights from the ViT encoder are transferred. ViT = Vision Transformer; FC = fully connected layer; CN = Cognitively normal; AD = Alzheimer's disease

we pre-train the model with separate MRI data unrelated to Alzheimer's disease or dementia, for their wide availability and abundance among public datasets.

In this paper, we leverage Masked Autoencoders (MAE) [9] as a self-supervised pre-training technique, due to its success with ViT. At first, an input 3D MRI image is divided into non-overlapping patches with a fixed, sine-cosine 3D positional embedding. Then, we follow the original implementation [9]: a subset of the patches is randomly masked and reconstructed by a lightweight decoder that takes as input the encoder representations of the visible subset of patches as well as the masked tokens, as shown in Fig 1. Similar to [9], we use an asymmetrical architecture for the decoder with 8 layers, 16 heads, but with an embedding hidden dimension of 576 due to 3D data dimensionality. This decoder architecture has a total of 36.9 million parameters.

Fine-tuning. The pre-trained weights of the encoder are transferred for fine-tuning, while the fully connected classification head is randomly initialized. The absolute sine-cosine positional embeddings are also transferred as initialization weights for learnable positional embeddings. Then, the training of the model was conducted in a supervised fashion using the labeled data.

We test MAE-based pre-training with three different masking ratios: 25%, 50%, and 75%. Furthermore, we explore the effect of the pre-training data size on downstream performance as well as the combination of several non-homogeneous pre-training datasets.

Fine-tuning with different amounts of data. The transformers-based models have been shown to excel in downstream tasks with few samples after pre-training [4]. We similarly investigate it in brain imaging and further study the generalizability of the ViT model under various low-data settings, that is, we fine-tune with different fractions of labeled training data – 10%, 20%, 40%, 60%, 80%.

Table 1. Dataset details. $P_m(\%)$ indicates the share of the majority class. CN = Cognitively normal; AD = Alzheimer's disease

Dataset	Magnet strength	AD	CN	P_m (%)
BRATS 2023	3T	–	1251	–
IXI	1.5T/3T	–	581	–
OASIS-3	3T	–	625	–
ADNI1	1.5T	192	229	54.4
ADNI2	3T	159	201	55.8

Hyperparameters ablation. In fine-tuning, we investigate the effect of data augmentation as well as a group of regularization methods of dropout, attention dropout, and drop path [11]. Following [8], our data augmentation strategy incorporates a set of 3D medical data augmentations, which includes random affine, random flipping, random rotation for 90 degrees, random scaling, and a random shift of intensity, all with a probability of 0.2.

Knowledge Distillation through attention, which was introduced in ViT with an additional distillation token by Touvron *et al.* [24], underscored the significance of knowledge distillation [10] in enhancing transformer architecture efficiency without supervised pre-training on the huge amount of external data. The distillation token is assumed to interact with all other self-attention embeddings and train with the distillation loss function [24], thus learning from a teacher model's predictions simultaneously. We similarly included this type of knowledge distillation in our ablation study.

2.2 Datasets

Pre-training datasets. For pre-training based on MAE, we utilized three different, public, and non-homogeneous T1-weighted structural MRI datasets: BRATS 2023 [2], [3], [20], IXI[1], and OASIS-3 [15]. Note that BRATS 2023 and IXI datasets are unrelated to dementia and Alzheimer's disease, while OASIS-3 includes images of individuals with various stages of cognitive decline, but we utilized images of only cognitively normal cases. We did brain extraction of all images of datasets using HD-BET [12]. More information on datasets is provided in Table 1.

Fine-tuning datasets. For AD classification experiments we collected two T1-weighted structural MRI datasets from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database[2] [13], namely baseline collections of ADNI1 and ADNI2. Each dataset contains structural MRI images of Alzheimer's disease patients and cognitively normal adults. Similar to pre-training datasets, we only

[1] https://brain-development.org/ixi-dataset

[2] https://adni.loni.usc.edu

Table 2. Cross-validation accuracies (%) for training from scratch and fine-tuning of MAE pre-training across different masking ratios (25%, 50%, and 75%). Pre-training significantly enhances accuracy. Best results are in **bold**.

		ADNI1	**ADNI2**
Training from scratch		66.0 ± 0.86	69.3 ± 1.81
MAE-based fine-tuning	25%	74.5 ± 1.36	74.0 ± 1.67
	50%	79.5 ± 1.09	79.2 ± 0.48
	75%	$\mathbf{79.6} \pm 0.82$	$\mathbf{81.9} \pm 2.17$

perform brain extraction to pre-process the data minimally. More information on datasets is also provided in Table 1.

2.3 Experimental setup

The default experimental setting is as follows. In all experiments, brain MRI images are transformed in the following order: images are first resampled to the same voxel spacing ($1.75 \times 1.75 \times 1.75$), followed by foreground crop, resizing ($128 \times 128 \times 128$), and intensity normalization.

Following He *et al.* [9], we pre-train the model using the AdamW optimizer [19] at an initial learning rate of 10^{-4}, and employ a half-cycle cosine scheduler with a 40-epoch linear warmup. We pre-train for 1000 epochs with a batch size of 32. On top of the abovementioned transformations, we only apply random spatial cropping.

In fine-tuning, we used an optimizer with an initial learning rate of 10^{-5} and a cosine annealing scheduler. We train using cross-entropy loss for 50 epochs with a batch size of 4. To determine a more accurate estimate of the performance in a low-data regime, we run each fine-tuning experiment as a stratified 4-fold cross-validation and use the best validation epoch to calculate the cross-validation accuracy. We repeat all experiments three times and report the average.

All experiments are implemented using PyTorch and data augmentations are accessed from MONAI [5]. The GPU configuration consisted of NVIDIA Titan RTX with 24 GB of VRAM.

3 Results and Discussion

In this section, we discuss experimental results and our key findings. Unless otherwise specified, reported results are for the pre-training with a masking ratio of 75% and using all combined pre-training datasets. Also, all results are with the best set of training methods found from our ablation study in Table 3.

3.1 Pre-training findings

Pre-training gives a considerable boost to the accuracy. Table 2 compares the performance of fine-tuning with training from scratch. In general, we

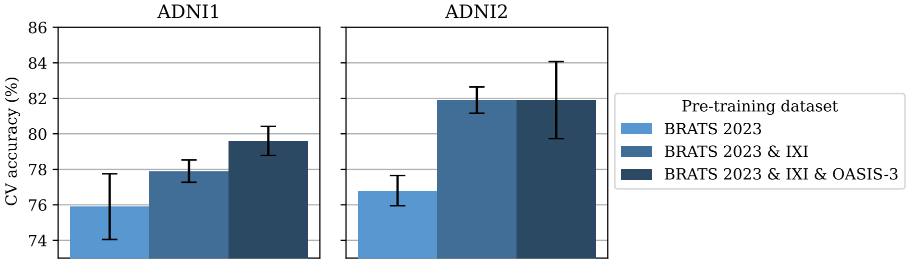

Fig. 2. Impact of the pre-training dataset size on fine-tuning. More pre-training data leads to better fine-tuning. Pre-training was performed with a 75% masking ratio.

observe that *fine-tuning the pre-trained weights enhances the accuracy of the ViT model by a considerable margin* in both AD classification datasets. When tested across different masking ratios, 75% exhibited the highest performance increase of up to 13.6% in ADNI1 and 12.6% in ADNI2. This is in line with the findings of the original implementation of MAE in computer vision [9], and other related works in medical imaging [26], [6]. Nevertheless, we emphasize that, in our experiments, we perform pre-training with the data unrelated to Alzheimer's disease or dementia. Thus, our results demonstrate the transferability of pre-trained features of ViT across different domains as well as its ability to boost performance in downstream applications.

Pre-training data size is crucial & combining non-homogeneous pre-training datasets is effective. Since we conducted pre-training with three various datasets, we also investigated the impact of pre-training data size on further fine-tuning accuracy. Fig. 2 illustrates that the pre-training data size is proportional to the fine-tuning accuracy, implying that *more pre-training data yields improved fine-tuning.* We also observe that *combining different non-homogeneous datasets from different sources is effective* in boosting the accuracy. Despite having a mix of magnetic strengths (Table 1) in the pre-training data, both ADNI1 and ADNI2 benefited with increasing accuracy. We conjecture that further experiments on pre-training with data augmentations would possibly result in improved performance, and leave it for future research.

Pre-training allows to succeed with fewer labeled data even under extreme low-data settings. Fig. 3 represents training with different fractions of labeled data ranging from 10% to 100%. Generally, we observe that *fine-tuning with as few as* 20% *of training set produces the model more accurate than training from scratch with* 100% *data.* Note that 20% of labeled data corresponds to roughly 60 samples from both datasets. As a consequence, we conclude that *the ViT model enormously benefits from pre-training when trained under low-data scenarios in brain imaging.* Self-supervised pre-training with the available data,

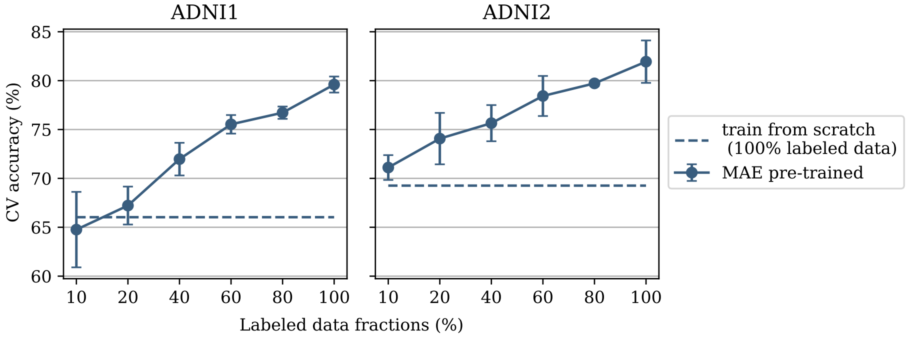

Fig. 3. Training with different fractions of labeled data. While pre-training is essential, labeled data remains critical as well.

which is unrelated to the downstream task, could therefore be advantageous in developing practical ViT models with small fine-tuning datasets. Notwithstanding, we note that the amount of labeled data remains crucial for further improvements.

3.2 Ablation study

Table 3 illustrates the ablation study with different training methods in order to find the most optimal training ingredients for fine-tuning the pre-trained ViT model in a low-data regime.

One *important component* for successful training is data augmentations, which elevated the accuracy by up to 3.4%. However, we *did not observe significant improvements* with regularization: including both stochastic drop path and attention dropout improve the performance, meanwhile, dropout had an apparent effect on the fine-tuning accuracy with the performance drop in both datasets ADNI1 and ADNI2. With a similar observation, Steiner *et al.* [23] concluded that regularization may hurt the performance as training data gets large. In our case, ADNI1 originally includes more samples than ADNI2 (Table 1) and had more performance drop, thus corroborating the conclusion.

Knowledge distillation through attention with a distillation token was another subject of ablation. We employed a 3D Resnet-152 network from the MONAI library as a teacher model. Similar to the above experimental setup, this model was trained separately for each of the three seeds[3]. Our experimentation, detailed in Table 3, first involved applying knowledge distillation to a randomly initialized model. Later, we explored the effects of applying distillation after the pre-training. Notably, *pre-training yielded a performance boost*

[3] Teacher network has average accuracy of 84.63% and 82.51% for ADNI2 and ADNI1 respectively.

Table 3. Ablation study on different training approaches for fine-tuning experiments.

Ablation ↓	**Data aug.**	**Drop path**	**Attn dropout**	**Dropout**	**Distill. token**	**ADNI1**	**ADNI2**
None (default)	✓	✓	✓	✗	✗	79.6 ±0.82	81.9 ±2.17
Data aug.	✗	✓	✓	✗	✗	77.4 ±2.06	78.0 ±1.15
Regularization	✓	✗	✓	✗	✗	78.9 ±1.44	81.4 ±1.44
	✓	✓	✗	✗	✗	78.2 ±0.90	80.7 ±2.33
	✓	✓	✓	✓	✗	78.7 ±1.07	80.0 ±1.69
Dist. w/o pre-train.	✓	✓	✓	✗	✓	67.7 ±0.62	69.8 ±2.52
Dist. with pre-train.	✓	✓	✓	✗	✓	79.7 ±1.19	82.0 ±1.37

compared to training from scratch and outperformed our model without distillation by 0.1%. However, our findings indicate that the primary performance improvement comes from the pre-training, showing its superiority as an effective method of enhancing model performance. Additionally, the knowledge distillation training requires increased GPU memory resulting from the utilization of a 3D convolutional teacher model.

4 Conclusion

We investigated the optimal training for the Vision Transformer (ViT) model in brain imaging, namely for Alzheimer's disease classification, in a low-data regime by leveraging various training strategies ranging from self-supervised pre-training to selecting an optimal set of training methods, including data augmentations, and regularizations. We demonstrated that pre-training immensely contributes to the increase in fine-tuning accuracy, even when trained under extreme low-data scenarios. We conducted all pre-training on the data unrelated to the downstream application, showing the generalizability of pre-trained features for fine-tuning disease classification. Additionally, while we showed that pre-training data size remains critical, we also confirmed that it is not only possible but also beneficial to combine different non-homogeneous datasets for pre-training. Finally, we presented an optimal training recipe which is useful in further boosting the fine-tuning accuracy with the ViT. We believe that this work will contribute to the development of optimal models not only for brain imaging but also for other medical imaging applications with limited data and computational resources.

As for limitations, we note the use of a single pre-training method and evaluation of a single task. Therefore we consider our future work to experiment with contrastive learning-based pre-training methods and extend our evaluations on more tasks, such as for example classifying mild cognitive impairment. Our future work will also continue with more exhaustive experiments on other Vision Transformer architectures and training methods.

Acknowledgments. This work was supported by the Engineering Research Center of Excellence (ERC) Program supported by National Research Foundation (NRF), Korean Ministry of Science & ICT (MSIT) (Grant No. NRF-2017R1A5A101470823).

Disclosure of Interests. The authors have no competing interests.

References

1. Arizumi, N.: Studying inductive biases in image classification task (2022), https://arxiv.org/abs/2210.17141
2. Baid, U., et al.: The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification (2021), https://arxiv.org/abs/2107.02314
3. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features. Scientific Data **4**(1) (Sep 2017)
4. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language Models are Few-Shot Learners (2020), https://arxiv.org/abs/2005.14165
5. Cardoso, M.J., Li, W., Brown, R., Ma, N., Kerfoot, E., Wang, Y., Murrey, B., Myronenko, A., Zhao, C., Yang, D., Nath, V., He, Y., Xu, Z., Hatamizadeh, A., Myronenko, A., Zhu, W., Liu, Y., Zheng, M., Tang, Y., Yang, I., Zephyr, M., Hashemian, B., Alle, S., Darestani, M.Z., Budd, C., Modat, M., Vercauteren, T., Wang, G., Li, Y., Hu, Y., Fu, Y., Gorman, B., Johnson, H., Genereaux, B., Erdal, B.S., Gupta, V., Diaz-Pinto, A., Dourson, A., Maier-Hein, L., Jaeger, P.F., Baumgartner, M., Kalpathy-Cramer, J., Flores, M., Kirby, J., Cooper, L.A.D., Roth, H.R., Xu, D., Bericat, D., Floca, R., Zhou, S.K., Shuaib, H., Farahani, K., Maier-Hein, K.H., Aylward, S., Dogra, P., Ourselin, S., Feng, A.: MONAI: An open-source framework for deep learning in healthcare (2022), https://arxiv.org/abs/2211.02701
6. Chen, Z., Agarwal, D., Aggarwal, K., Safta, W., Hirawat, S., Sethuraman, V., Balan, M.M., Brown, K.: Masked Image Modeling Advances 3D Medical Image Analysis (2022), https://arxiv.org/abs/2204.11716
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ICLR (2021)
8. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H.R., Xu, D.: UNETR: Transformers for 3D medical image segmentation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 574–584 (2022)

9. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16000–16009 (2022)
10. Hinton, G., Vinyals, O., Dean, J.: Distilling the Knowledge in a Neural Network (2015), https://arxiv.org/abs/1503.02531
11. Huang, G., Sun, Y., Liu, Z., Sedra, D., Weinberger, K.: Deep Networks with Stochastic Depth (2016), https://arxiv.org/abs/1603.09382
12. Isensee, F., Schell, M., Pflueger, I., Brugnara, G., Bonekamp, D., Neuberger, U., Wick, A., Schlemmer, H.P., Heiland, S., Wick, W., Bendszus, M., Maier-Hein, K.H., Kickingereder, P.: Automated brain extraction of multisequence MRI using artificial neural networks. Human Brain Mapping **40**(17), 4952–4964 (aug 2019)
13. Jack, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., Whitwell, J.L., Ward, C., Dale, A.M., Felmlee, J.P., Gunter, J.L., Hill, D.L., Killiany, R., Schuff, N., Fox-Bosetti, S., Lin, C., Studholme, C., DeCarli, C.S., Krueger, G., Ward, H.A., Metzger, G.J., Scott, K.T., Mallozzi, R., Blezek, D., Levy, J., Debbins, J.P., Fleisher, A.S., Albert, M., Green, R., Bartzokis, G., Glover, G., Mugler, J., and, M.W.W.: The alzheimer's disease neuroimaging initiative (ADNI): MRI methods. Journal of Magnetic Resonance Imaging **27**(4), 685–691 (2008)
14. Jang, J., Hwang, D.: M3T: three-dimensional Medical image classifier using Multi-plane and Multi-slice Transformer. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20686–20697. IEEE, New Orleans, LA, USA (Jun 2022)
15. LaMontagne, P.J., Benzinger, T.L., Morris, J.C., Keefe, S., Hornbeck, R., Xiong, C., Grant, E., Hassenstab, J., Moulder, K., Vlassenko, A.G., Raichle, M.E., Cruchaga, C., Marcus, D.: OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer Disease. medRxiv (2019)
16. Lee, S.H., Lee, S., Song, B.C.: Vision Transformer for Small-Size Datasets (2021), https://arxiv.org/abs/2112.13492
17. Li, C., Cui, Y., Luo, N., Liu, Y., Bourgeat, P., Fripp, J., Jiang, T.: Trans-ResNet: Integrating Transformers and CNNs for Alzheimer's disease classification. In: 2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI). pp. 1–5 (Mar 2022), iSSN: 1945-8452
18. Liu, Y., Sangineto, E., Bi, W., Sebe, N., Lepri, B., Nadai, M.D.: Efficient Training of Visual Transformers with Small Datasets (2021), https://arxiv.org/abs/2106.03746
19. Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization (2019), https://arxiv.org/abs/1711.05101
20. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., et al.: The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE Transactions on Medical Imaging **34**(10), 1993–2024 (Oct 2015)
21. Nirthika, R., Manivannan, S., Ramanan, A., Wang, R.: Pooling in convolutional neural networks for medical image analysis: a survey and an empirical study. Neural Computing and Applications **34**(7), 5321–5347 (feb 2022)
22. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning Transferable Visual Models From Natural Language Supervision (2021), https://arxiv.org/abs/2103.00020
23. Steiner, A., Kolesnikov, A., Zhai, X., Wightman, R., Uszkoreit, J., Beyer, L.: How to train your vit? data, augmentation, and regularization in vision transformers (2022), https://arxiv.org/abs/2106.10270

24. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International conference on machine learning. pp. 10347–10357. PMLR (2021)
25. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention Is All You Need (2023), https://arxiv.org/abs/1706.03762
26. Zhou, L., Liu, H., Bae, J., He, J., Samaras, D., Prasanna, P.: Self Pre-training with Masked Autoencoders for Medical Image Classification and Segmentation (2023), https://arxiv.org/abs/2203.05573