Parameter Efficient Fine-tuning of Transformer-based Masked Autoencoder Enhances Resource Constrained Neuroimage Analysis

Nikhil J. Dhinagar[1], Saket S. Ozarkar[1], Ketaki U. Buwa[1], Sophia I. Thomopoulos[1],
Conor Owens-Walton[1], Emily Laltoo[1], Chirag Jagad[1], Yao-Liang Chen[2], Philip Cook[3],
Corey McMillan[3], Chih-Chien Tsai[4], J-J Wang[5], Yih-Ru Wu[6], Paul M. Thompson[1]
[1]Imaging Genetics Center, Mark & Mary Stevens Neuroimaging & Informatics Institute,
Keck School of Medicine, University of Southern California, Los Angeles, CA, USA
[2]Department of Diagnostic Radiology, Chang Gung Memorial Hospital, Keelung, Taiwan
[3]Department of Neurology, Perelman School of Medicine,
University of Pennsylvania, Philadelphia, PA, USA
[4]Healthy Aging Research Center, Chang Gung University, Taoyuan, Taiwan
[5]Department of Medical Imaging and Radiological Sciences, Chang Gung University, Taoyuan, Taiwan
[6]Department of Neurology, Chang Gung Memorial Hospital, Linkou, Taoyuan, Taiwan

ABSTRACT

Recent innovations in artificial intelligence (AI) have increasingly focused on large-scale foundational models that are more general purpose in contrast to conventional models trained to perform specialized tasks. Transformer-based architectures have become the standard backbone in foundation models across data modalities (image, text, audio, video). There has been a keen interest in applying parameter-efficient fine-tuning (PEFT) methods to adapt these models to specialized downstream tasks in language and vision. These methods are particularly essential for medical image analysis where the limited availability of training data could lead to overfitting. In this work, we evaluated different types of PEFT methods on pre-trained vision transformers relative to typical training approaches, such as full fine-tuning and training from scratch. We used a transformer-based masked autoencoder (MAE) framework, to pretrain a vision encoder on T1-weighted (T1-w) brain MRIs. The pretrained vision transformers were then fine-tuned using different PEFT methods that reduced the trainable model parameters to as few as 0.04% of the original model size. Our study shows that: 1. PEFT methods were competitive with or outperformed the reference full fine-tuning approach and outperformed training from scratch, with only a fraction of the trainable parameters; 2. PEFT methods with a 32% reduction in model size boosted Alzheimer's disease (AD) classification by 3% relative to full fine-tuning and 11% relative to a 3D CNN, with only 258 training scans; and 3. PEFT methods performed well on diverse neuroimaging tasks including AD and Parkinson's disease (PD) classification, and "brain-age" prediction based on T1-w MRI datasets - a standard benchmark for deep learning models in neuroimaging; 4. smaller model sizes were competitive with larger models in test performance. Our results show the value of adapting foundation models to neuroimaging tasks efficiently and effectively in contrast to training stand-alone special purpose models.

Keywords: Parameter-efficient fine-tuning, masked autoencoder, vision transformer, neuroimaging, MRI

1. INTRODUCTION

Recent advances in AI have greatly advanced medical image analysis. AI techniques can be used in clinical practice as well as for neuroscience and radiology research. Many promising studies have trained deep learning methods on brain magnetic resonance imaging (MRI) data collected worldwide to perform tasks such as differential diagnosis or prognosis in dementia. Some key challenges in adapting AI to medical tasks include working with limited labeled training data,

This work was supported by the U.S. National Institutes of Health, under NIH grant U01 AG068057.

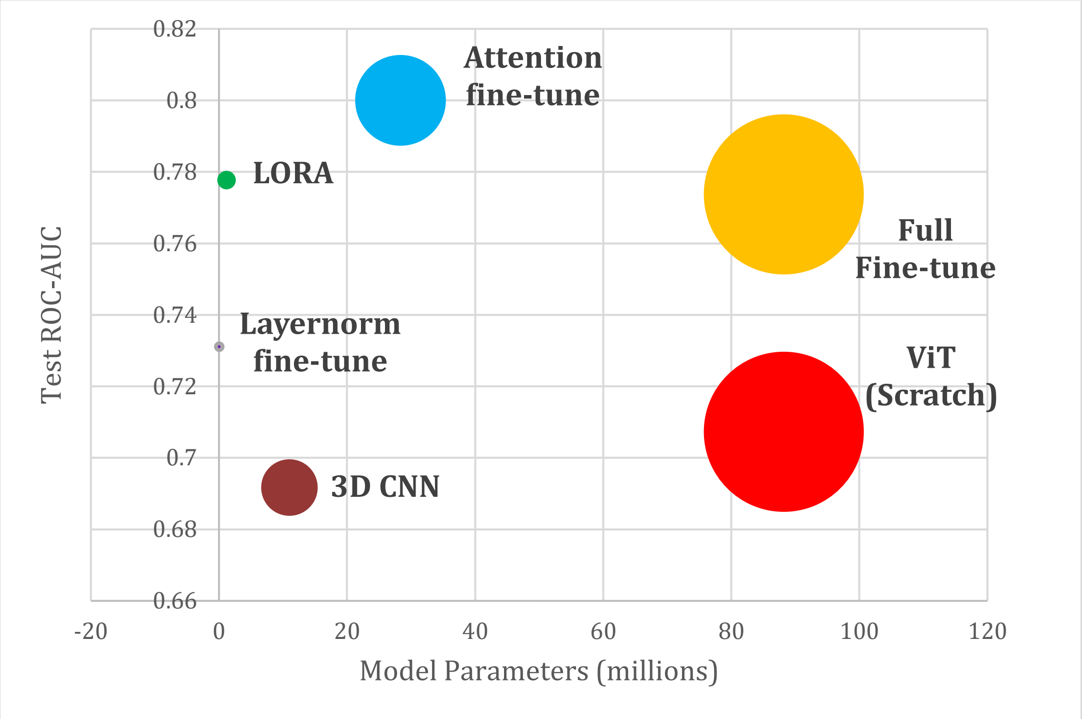

Figure 1. Test performance and vision transformer (ViT-B) model parameter tradeoff using different PEFT methods relative to full fine-tuning for AD classification - with *only 258 training MRIs* from the ADNI dataset. Larger bubble sizes indicate a greater number of model parameters to be tuned for the downstream task. Compared to very large models (on the right), attention fine-tuned and LORA models maintain high accuracy with far fewer parameters (top left).

privacy requirements when handling medical data, and heterogeneity of data collected at different sites and scanners. The 'transfer learning' paradigm, which pre-trains AI models first and then fine-tunes them for specific tasks, has helped to overcome the lack of large datasets in medical applications. Many pre-training approaches have been proposed and tested, including methods based on (semi-) supervised learning and contrastive learning.[1] [2] [3] [4] Masked autoencoders provide an effective way to pretrain vision transformers. [5] [6]

The pre-trained models are then adapted to a downstream task by fully fine-tuning all the neural network layers or some of the layers. As large-scale foundation models have emerged, [7] methods are needed to efficiently fine-tune them. The use of large vision transformer-based models is challenging, as they are prone to overfitting with smaller datasets. Parameter-efficient fine-tuning (PEFT) methods have seen increased interest in the context of language models before being adopted for computer vision applications as well. PEFT methods only fine-tune a small fraction of the model parameters and freeze the remainder. The tunable model parameters in the PEFT methods are usually modified through *additive* or *selective* approaches applied to different sections and mechanisms of the model. PEFT methods also make it possible to fine-tune larger models on limited computational consumer hardware. PEFT methods include adaptors, [8] [9] low-rank approximation (LORA), [10] prompt tuning, [11] and attention tuning.[12] Some recent work shows the benefits of PEFT methods for medical applications. [13] [14] For some tasks, PEFT may perform as well as full fine-tuning, at a fraction of the computational cost. In this work, we evaluated the performance of different *additive* and *selective* PEFT methods with the vision transformer. We tested the effect of PEFT methods on the sample size requirements for the downstream fine-tuning dataset. We also demonstrated that these methods can be used to adapt the vision encoders to diverse neuroimaging tasks and their corresponding datasets.

Table 1. Model Parameter count (in millions) using parameter-efficient fine-tuning methods with the ViT-B and ViT-S.

Backbone	Scratch	Full fine-tune	Attention fine-tune	LORA	Layernorm fine-tune
ViT-B	88,204,034	88,204,034 (100%)	28,349,954 (32%)	1,181,308 (1.3%)	38,400 (0.04%)
ViT-S	12,220,800	12,220,800 (100%)	3,548,930 (29%)	295,804 (2.4%)	10,754 (0.09%)

2. METHODS

In this work, we used the vision transformer (ViT) architecture, [15] specifically, the ViT-Base (ViT-B), [16] and its smaller variant, ViT-Small (ViT-S). Each of these models is based on the transformer architecture, currently used across diverse data modalities (image, video, audio, text) ubiquitously. The ViT-B encoder has 88 million trainable parameters, and the ViT-S encoder has 12 million trainable parameters. We pretrained the ViT encoders on the large-scale UK Biobank (UKBB) brain MRI datasets using the MAE framework. [5] [6] 75% of patches generated from the T1-w MRIs were masked initially and a lightweight decoder was used to predict the masked patches.

Given the state-of-the art nature of this framework and relatively large, diverse pre-training data, this foundation model is expected to *generalize* well to downstream tasks. In the next section, we describe PEFT methods that enabled this foundation model to be further *adapted* to achieve the best performance possible on specialized tasks and their datasets. In this paper, we evaluated two main types of PEFT methods, shown in **Table 1**: *additive* methods that require inclusion of new parameters and *selective* methods that optimize only a subset of the model's parameters. In both methods, the underlying original model parameters are frozen and only the new or subset parameters are fine-tuned. As an *additive* method we tested low-rank adaptation (LORA), [10] [17] where the pre-trained attention weight matrix is frozen and trainable rank decomposition matrices are injected in the transformer encoder layers. The *selective* methods consisted of two approaches, attention fine-tune, and layer normalization fine-tune, where only the attention and layer norm layers are respectively fine-tuned. As a baseline, we compared our PEFT methods with the encoders trained from scratch (without any pre-training) and full fine-tuning involving all the original trainable model parameters. We also used a 3D DenseNet121 CNN with 11 million parameters as an additional reference for our experiments.

3. EXPERIMENTAL SETUP

3.1 Data

In line with similar studies, [2] all 3D T1-weighted brain MRI scans were pre-processed via standard steps for neuroimaging analyses, including: nonparametric intensity normalization (N4 bias field correction), [18] 'skull-stripping', linear registration to a template with 9 degrees of freedom, and isotropic resampling of voxels to 2-mm resolution. The input spatial dimension of the MRIs was 91x109x91. All images were *z*-transformed (setting each image's mean and SD to a standard value) to stabilize model training. The T1-w scans were re-sized to 80 voxels across all dimensions before model training. In this work, we also used 3D T1-w brain MRI scans from the UK Biobank [19] dataset to pretrain the vision encoders using the MAE framework (see Table 2 for details). We used data from the publicly available ADNI dataset [20] for the downstream tasks of Alzheimer's disease (AD) classification and brain age prediction. We used two additional datasets for Parkinson's disease (PD) classification. We also used the OASIS [21] and UPenn datasets as out-of-distribution datasets for zero-shot testing for each experiment. For the brain age experiments, we used a subset of ADNI with only MRIs from healthy control participants. The datasets used in this study are summarized in **Table 2**.

Table 2. Summary of data utilized in this work for pretraining and fine-tuning.

Dataset	N	#Subjects	Age (years)	Sex (F/M)
UKB	38,703	38,703	44.6-82.8	20,216F/18,487M
ADNI	4,098	1,188	55.7-92.8	556F/632M
OASIS	600	600	43.5-97.0	341F/359M
Taiwan	467	467	20.0-80.0	236F/231M
UPenn	164	164	50.0-86.0	63F/101M

3.2 Model Training and Testing

We performed a random search to select hyperparameter values including the batch size, learning rate, learning rate scheduler warmup epochs, weight decay, and dropout for pretraining and finetuning. The vision transformers in the MAE had the following specifications, ViT-B - Encoder: 12 layers, 12 attention heads and Decoder: 8 layers, 12 attention heads; and ViT-S – Encoder: 6 layers, 8 attention heads and Decoder: 8 attention heads, 4 layers. A patch size of 16 was used for the ViTs. The ViTs were pre-trained for 1,000 epochs and fine-tuned for 30 to 50 epochs with an early stopping of 10 epochs. All models used the AdamW optimizer and minimized the mean squared error (MSE) loss for pre-training, the L1, binary cross entropy loss functions for regression and classification fine-tuning tasks respectively. We evaluated our models with the receiver-operator characteristic curve-area under the curve (ROC-AUC), and mean absolute error (MAE) for classification and regression, respectively. In our results, we present an average over multiple runs with three different seeds. We conducted zero-shot testing with an independent dataset for each of the fine-tuning tasks.

4. RESULTS

The ViTs were trained from scratch and fully fine-tuned as a baseline. We evaluated the PEFT methods relative to training from scratch, with full fine-tuning -- as well as the 3D DenseNet121 CNN. We investigated the use of PEFT methods for different neuroimaging tasks and datasets as presented in **Tables 3 and 4**. **Figure 1** Illustrates the tradeoff between test performance and the number of model parameters; **Figures 2 and 3** shows the test ROC-AUC versus training data sizes using ADNI, where the different curves represent the various PEFT methods.

5. DISCUSSION

1. Test performance for PEFT is competitive with or outperforms full fine-tuning, but at a reduced cost.

The PEFT methods reduced the number of trainable parameters significantly compared to full fine-tuning i.e., 30% to 0.04% of the total model size. For AD classification, attention tuning matched the full fine-tune performance on the ADNI test set when trained with 100% of the training data but with a reduced model size – i.e., only using 32% of the original number of model parameters. With LORA (1.3% model parameters) and layernorm fine-tune (0.04% model parameters), these methods give a considerable reduction in model size. We recorded only a slight drop in the test ROC-AUC relative to full fine-tune. Another key takeaway was that all of the PEFT methods were comparable in performance with the full fine-tune on the zero-shot OASIS test set. For brain age prediction, all the PEFT methods outperformed both training from scratch and full fine-tuning. For ViT-B, attention fine-tuning was the best-performing PEFT method achieving a reduction in MAE by 0.42 years. For PD classification, all PEFT methods showed only a slight drop in performance compared to full fine-tune but always outperformed training from scratch. LORA and Layernorm fine-tune outperformed full fine-tune on the zero-shot UPenn dataset. For both the brain age prediction and PD classification tasks, PEFT outperforms the 3D DenseNet121 CNN.

Table 3. Application of PEFT methods using the ViT-B for diverse tasks, relative to full fine-tuning.

Backbone		**ViT-B**					**DenseNet121 CNN**
Task	**PEFT**	**Scratch**	**Full fine-tune**	**Attention fine-tune**	**LORA**	**Layernorm fine-tune**	**Scratch**
	Dataset						
Alzheimer's Disease (AD) Classification, Test ROC-AUC	ADNI	0.8531 (0.0304)	**0.8619 (0.0241)**	**0.8613 (0.0094)**	0.8153 (0.0500)	0.8077 (0.0269)	0.8865 (0.013)
	OASIS (zero-shot)	0.8184 (0.0040)	0.8353 (0.0103)	0.8311 (0.0075)	0.8295 (0.0162)	0.8263 (0.0107)	0.8371 (0.0080)
Brain Age Prediction, Test MAE	ADNI	4.9140 (0.3700)	**4.5076 (0.1530)**	**4.0860 (0.1400)**	**4.4836 (0.1900)**	**4.2463 (0.5070)**	4.4255 (0.0212)
	OASIS (zero-shot)	9.6226 (1.0190)	8.0880 (0.9810)	6.7593 (0.4060)	8.4426 (0.3710)	7.6873 (1.4000)	7.8484 (0.2870)
Parkinson's Disease (PD) classification, Test ROC-AUC	Taiwan	0.5424 (0.0927)	**0.6521 (0.0400)**	**0.6218 (0.1073)**	**0.6200 (0.0109)**	0.5962 (0.0281)	0.5891 (0.0263)
	UPenn (zero-shot)	0.4791 (0.0873)	0.5457 (0.0325)	0.5414 (0.1170)	0.5710 (0.0019)	0.5687 (0.0648)	0.6100 (0.0393)

Table 4. Application of PEFT methods using the ViT-S for AD classification, relative to full fine-tuning.

Backbone		**ViT-S**				
Task	**PEFT**	**Scratch**	**Full fine-tune**	**Attention fine-tune**	**LORA**	**Layernorm fine-tune**
	Dataset					
Alzheimer's Disease (AD) Classification, Test ROC-AUC	ADNI	0.8426 (0.0189)	**0.8766 (0.0092)**	**0.8633 (0.0101)**	0.7886 (0.0267)	0.7987 (0.0648)
	OASIS (zero-shot)	0.8316 (0.0053)	0.8442 (0.0087)	0.8306 (0.0163)	0.8198 (0.0078)	0.8122 (0.0417)

2. PEFT methods boosted performance in limited data settings.

As shown in **Figure 1,** for ViT-B, the attention fine-tune and LORA methods outperformed (+0.004% to +3% test ROC-AUC) full fine-tuning when using 10% (258 scans) of the ADNI training data set for AD classification. We illustrate the overall effect of the downstream training dataset size (10% to 100%) on the PEFT test performance in **Figure 1.** All PEFT methods tested outperformed training from scratch by 3% to 10% in terms of the test ROC-AUC for AD classification. All the PEFT methods also outperformed the 3D DenseNet 121 CNN for AD classification when using only 258 training scans, including a performance boost of almost 11% with the attention fine-tune method. The PEFT provides the ViT with the capability to be competitive with the 3D DenseNet121 CNN, even with limited training data.

3. PEFT helped in adaptation to multiple downstream tasks and datasets.

We tested the effectiveness of the PEFT with multiple downstream neuroimaging tasks and multiple datasets as shown in **Table 3**. These tasks were chosen to provide a diverse set of challenges to evaluate the PEFT methods. The PEFT methods were competitive or outperformed full fine-tuning with a substantial reduction in the number of trainable parameters for AD, PD classification and brain age prediction. The PEFT methods also usually outperformed training the ViTs from scratch.

4. Smaller models are competitive with larger model performance

We tested the different PEFT methods as a function of the ViT model size and observed that the smaller ViT-S was competitive and sometimes outperformed the larger model in AD classification. This is likely due to the limited size of typical neuroimaging datasets; the approach helps with efficient fine-tuning of foundation models for new tasks. Like the larger ViT (ViT-B), the PEFT methods help boost ViT-S test-time performance and crucially in low-training data regimes as seen in **Figure 2**. Attention-tuning is the best performing PEFT method for ViT-S for AD classification similar to the ViT-B architecture, both relative to full-finetuning.

We plan to extend this work further with other PEFT methods, additional downstream tasks and other model architectures.

5. CONCLUSION

In this work, we evaluated multiple parameter-efficient fine-tuning methods for neuroimaging tasks and datasets. Our experiments demonstrated the benefit of parameter-efficient fine-tuning methods applied to pre-trained vision transformer encoders in resource constrained environments. The PEFT methods were competitive with or outperformed full fine-tuning with fewer model parameters. PEFT boosted performance in low training-data settings. We also show that the PEFT methods can be used to adapt pre-trained backbones to a range of tasks. Additionally, we tested multiple model sizes in a series of ablation experiments for each of the PEFT methods.

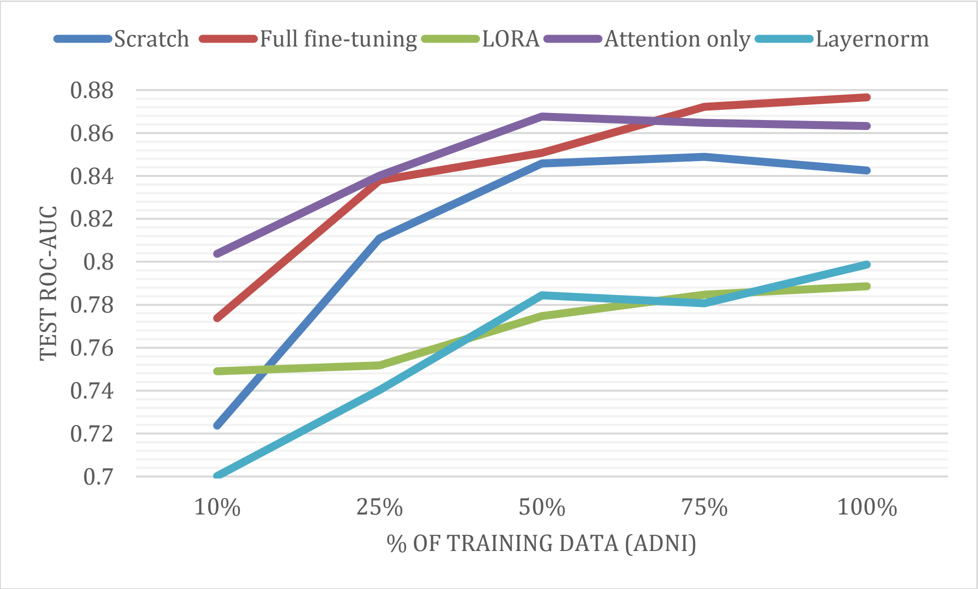

Figure 2. ViT-S: Ablation study shows the effect of the amount of downstream fine-tuning data vs the fine-tuning methods, Test AUC vs % training data for AD classification.

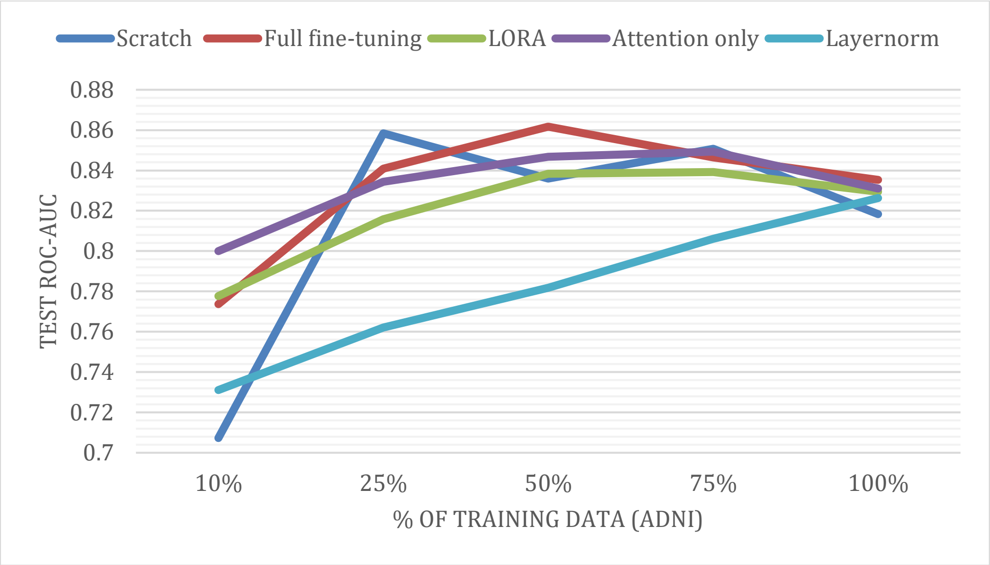

Figure 3. VIT-B: Ablation study shows the effect of the amount of downstream fine-tuning data vs the fine-tuning methods, Test AUC vs % training data for AD classification.

6. REFERENCES

[1] Dhinagar, N. J., Singh, A., Ozarkar, S. S., Buwa, K. U., Thomopoulos, S., Owens-Walton, C., Laltoo, E., Chen, Y.-L., Cook, P. A., McMillan, C., Tsai, C.-C., Wang, J.-J., Wu, Y.-R. and Thompson, P. M., “Video and synthetic MRI pre-training of 3D vision architectures for neuroimage analysis,” SPIE Med. Imaging (2024).

[2] Dhinagar, N. J., Thomopoulos, S. I., Rajagopalan, P., Stripelis, D., Ambite, J. L., Steeg, G. Ver and Thompson, P. M., “Evaluation of Transfer Learning Methods for Detecting Alzheimer’s Disease with Brain MRI,” SIPAIM (2022).

[3] Dufumier, B., Gori, P., Victor, J., Grigis, A., Wessa, M., Brambilla, P., Favre, P., Polosan, M., McDonald, C., Piguet, C. M., Phillips, M., Eyler, L. and Duchesnay, E., “Contrastive Learning with Continuous Proxy Meta-data for 3D MRI Classification,” MICCAI (2021).

[4] Chen, T., Kornblith, S., Norouzi, M. and Hinton, G., “A Simple Framework for Contrastive Learning of Visual Representations,” ICML, 1597–1607 (2020).

[5] Feichtenhofer, C. and Ai, M., “Masked Autoencoders As Spatiotemporal Learners” NeurIPS (2022).

[6] Zhou, L., Liu, H., Bae, J., He, J., Samaras, D. and Prasanna, P., “Self Pre-training with Masked Autoencoders for Medical Image Classification and Segmentation” IEEE ISBI (2022).

[7] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., et al., “On the Opportunities and Risks of Foundation Models,” Stanford HAI, Technical Report, 1–214 (2021).

[8] Chen, S., Ge, C., Tong, Z., Wang, J., Song, Y., Wang, J. and Luo, P., “AdaptFormer: Adapting Vision Transformers for Scalable Visual Recognition,” NeurIPS (2022).

[9] Houlsby, N., Giurgiu, A., Jastrzçbski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., Attariyan, M. and Gelly, S., “Parameter-efficient transfer learning for NLP,” 36th Int. Conf. Mach. Learn. ICML 2019 2019-June, 4944–4953 (2019).

[10] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L. and Chen, W., “LoRA: Low-Rank Adaptation of Large Language Models,” ICLR (2022).

[11] Lester, B., Al-Rfou, R. and Constant, N., “The Power of Scale for Parameter-Efficient Prompt Tuning,” EMNLP 2021 - 2021 Conf. Empir. Methods Nat. Lang. Process. Proc., 3045–3059 (2021).

[12] Shi, B., Gai, S., Darrell, T. and Wang, X., “TOAST: Transfer Learning via Attention Steering” arXiv (2023).

[13] Dutt, R., Ericsson, L., Sanchez, P., Tsaftaris, S. A. and Hospedales, T., “Parameter-Efficient Fine-Tuning for Medical Image Analysis: The Missed Opportunity” MIDL (2024).

[14] Wu, J., Fu, R., Fang, H., Liu, Y., Wang, Z., Xu, Y., Jin, Y. and Arbel, T., “Medical SAM Adapter: Adapting Segment Anything Model for Medical Image Segmentation,” arXiv (2023).

[15] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Houlsby, N., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” Int. Conf. Learn. Represent. (2021).

[16] Dhinagar, N. J., Thomopoulos, S. I., Laltoo, E. and Thompson, P. M., “Efficiently Training Vision Transformers on Structural MRI Scans for Alzheimer’s Disease Detection,” Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBC (2023).

[17] Zhu, Y., Shen, Z., Zhao, Z., Wang, S., Wang, X., Zhao, X., Shen, D. and Wang, Q., “MeLo: Low-rank Adaptation is Better than Fine-tuning for Medical Image Diagnosis” ISBI (2024).

[18] Tustison, N. J., Cook, P. A. and Gee, J. C., “N4ITK: Improved N3 Bias Correction,” IEEE Trans Med Imaging. **29**(6), 1310–1320 (2010).

[19] Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., Liu, B., Matthews, P., Ong, G., Pell, J., Silman, A., Young, A., Sprosen, T., Peakman, T. and Collins, R., “UK Biobank: An Open Access Resource for Identifying

the Causes of a Wide Range of Complex Diseases of Middle and Old Age," PLoS Med. **12**(3), 1–10 (2015).

[20] Weiner, M. W. and Veitch, D. P., "Introduction to Special Issue Overview of ADNI," Alzheimers Dement **11**(7), 730–733 (2015).

[21] LaMontagne, P. J., Benzinger, T. L., Morris, J. C., Keefe, S., Hornbeck, R., Xiong, C., Grant, E., Hassenstab, J., Krista, M., Vlassenko, A. G., Raichle, M. E., Cruchaga, C. and Marcus, D., "OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer Disease" medRxiv (2019).