Swin Transformer for MRI-based Brain Age Estimation

Pradyumna Rao
Dept. of Electrical and Computer Engineering
Rutgers University
pradyumna.rao@rutgers.edu

Manasa Mangipudi
Dept. of Electrical and Computer Engineering
Rutgers University
mm3754@scarletmail.rutgers.edu

Abstract

Estimating brain age can be critically important in clinical settings to help with the prognosis and diagnosis of neurodegenerative diseases such as Alzheimer's Disease and other related disorders such as traumatic brain injuries (TBI) because early detection and the ensuing treatment greatly improve patient quality of life. In settings such as neural imaging and deep learning, convolutional neural network (CNN) frameworks have long been used especially due to their reliability in performance and general ability to achieve good results even with small datasets. Recently, we have seen the vision transformer (ViT) emerge as an alternative to CNNs, achieving even better performance with the caveat that they require large training datasets. In this paper, we show that even with a small dataset (n=156), it is possible, using transfer learning and fine tuning, to achieve better results using ViT than a CNN. We implement a pretrained Swin Transformer and fine tune it on the Open Neuro dataset, obtaining 92.5% accuracy as opposed to 85% on the ResNet-18 framework. We share our code on GitHub at the following repo: https://github.com/roborao/Swin-Brain-Age-Estimation.

Index Terms

Vision Transformer, Swin, Brain Age Estimation, Deep Learning

I. Introduction

With the introduction of more powerful hardware, machine learning - and in particular, deep learning - has taken off. We are highly interested in utilizing these deep learning models in the domain of medical imaging. Medical imaging has been an ongoing and open area of research in the specific domains of tumor detection, EEG event detection, and neural activity detection and classification. Yet another interesting area of research is identifying brain age based on MRI scans. Being able to estimate patients' brain age, and therefore, their implicit brain health, could be a key development in being able to identify neurodegenerative diseases that are afflicting people, even before they become symptomatic. Physicians may also be able to use this information to then provide their patients with visual proof of localized neurodegeneration, allowing them to alter lifestyle habits, such as sleep or screen time, or seek treatment, if available.

In the space of machine vision, we have many tools available to use for feature detection, such as scale-invariant feature transform (SIFT) or Harris corner detection. However, with neuroimaging in particular, corner detection may not be effective, as there are many blob-like and curved features. In addition, detecting many keypoints in a brain scan may be a highly noisy process, and indeed, one that could lead to a high degree of visual clutter, leading to a lack of interpretability on the physician or technician's end. Ironically, while machine learning architectures such as neural networks have been criticized for a lack of interpretability, as they have a tendency to be black-box universal function approximators, they excel in classification problems, such as the one posed here. Historically, CNNs have been proposed for

image classification; they excel in image processing due to the convolution operator which slides across subsections of the image, extracting features and then downsampling them in the following pooling layer. CNNs are also known for their parallelizability and efficient performance. It is important to note that in the medical domain, prognoses and diagnoses are made with a certain degree of confidence, but using deep learning techniques we want to maximize this confidence level, as to not only provide the correct patient care regimen, but also peace of mind for the patient.

Recently, we have seen an explosion in the use of the relatively newer so-called Transformer architecture, which excel in NLP tasks. ViT was first introduced in the seminal paper, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" by Dosovitskiy et al [1], building off an encoder-only Transformer architecture and breaking the dominance of CNNs in image classification problems. Much more recently, the Swin Transformer achieved breakthrough performance by combining the convolution sliding window with the attention mechanism to create hierarchical feature maps [2]. The main difference between the vanilla ViT and Swin ViT architectures is shown below, in Fig. 1. The goal of using the transformer architecture for this problem, in general, is to discover if the model can be forced to pay attention to the defining characteristics of a developing, mature, and aging brain. One problem to note is the general lack of labeled data in medical imaging tasks; we aim to solve this problem using data augmentation and transfer learning, which Asiri et al. [3] show as effective approaches for neuroimaging. Indeed, in this paper, we show an improvement over CNNs by using transformers in the brain age classification using pretrained and fine-tuned Swin ViTs.

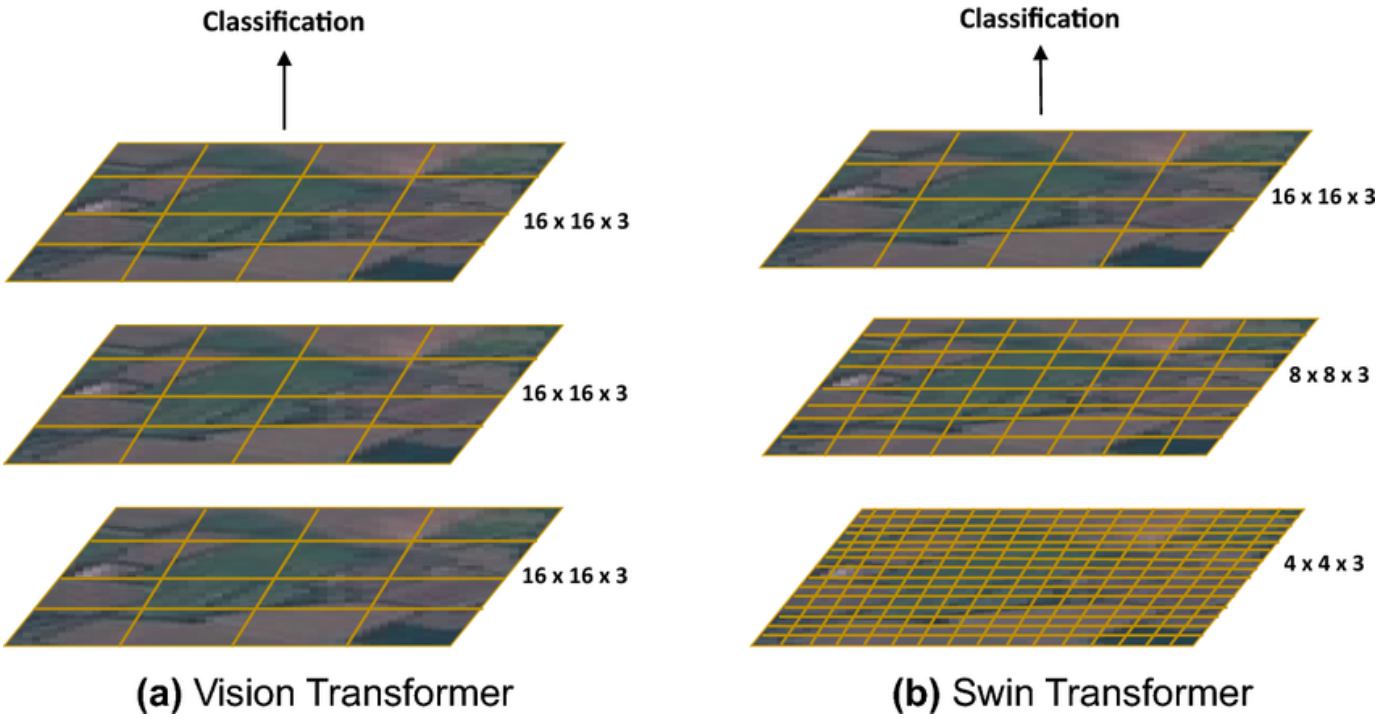

Fig. 1. ViT vs Swin sliding window mechanism

II. Related Works

Recent studies have demonstrated diverse approaches in brain MRI analysis through deep learning architectures. Wahlang et al. [4] incorporated demographic variables directly into their neural network structure, while Dinsdale et al. [5] utilized specialized CNNs for age-related pattern recognition. He et al. [6] applied LSTM networks to pediatric brain analysis, demonstrating effectiveness in capturing developmental patterns, and Illakiya and Karthik [7] combined attention mechanisms with Swin Transformer [2] for MCI classification.

Vision Transformers (ViT) [1] have emerged as powerful alternatives to CNNs in medical image analysis, offering superior performance in capturing long-range dependencies. While traditional ViTs process images by dividing them into fixed-size patches, Swin Transformers introduce a hierarchical feature representation with shifted windows, enabling more efficient processing of high-resolution medical images [8]. As discussed, the hierarchical design of Swin Transformers particularly addresses the limitations of standard ViTs in handling varying scales of features, making them especially suitable for complex brain MRI analysis where both local and global features are crucial for accurate diagnosis.

III. Dataset

For this paper, the dataset we used is publicly available from OpenNeuro [9]. Our dataset consists of MRI brain scans of subjects from 3 age groups, children of age 3-5, children of age 7-12 and adults, while they watched "Partly Cloudy," a Disney Pixar short film. Our dataset consists of MRI scan images of each subject, with the distribution being 65 subjects in age 3-5 category, 57 subjects in age 7-12 category and 33 subjects in adults category. In order to address the issue of small dataset, we increased the dataset size by performing data augmentation. As a part of data augmentation, we flipped all original images by 180 degrees, doubling the dataset size. A sample image of our processed data, after converting it from RGB mode to grayscale and resizing it to (224,224) size, is shown in Fig. 2

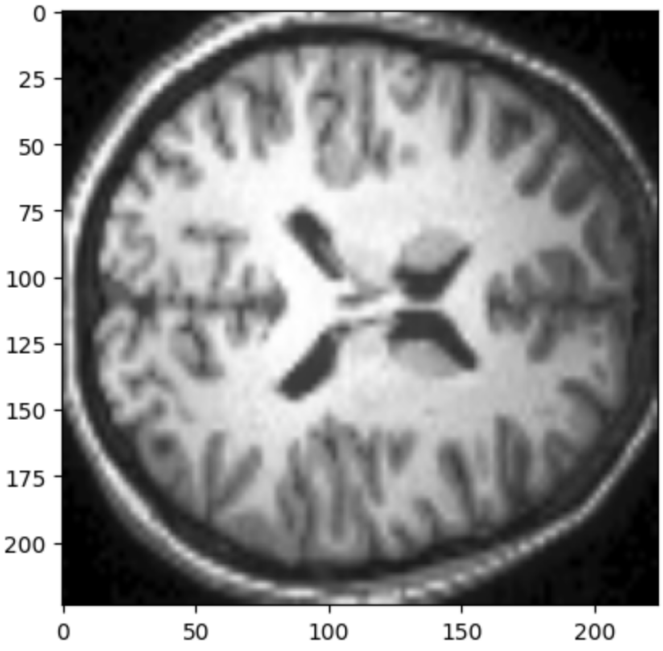

Fig. 2. MRI Scan Image of an Adult from dataset

IV. Experiments & Results

We will now conduct experiments using the ViT architecture to perform simple ternary classification between three groups, children ages 3-5, children ages 7-12, and adults using the above-mentioned OpenNeuro dataset. We train on a Google Tesla T4 GPU with 16 GB RAM. Using pretrained models for the following experiments is necessary because of ths relatively small sample size of the dataset. Thus, transfer learning and fine tuning is used to train the model for our specific dataset. The code used for our training is shared at the following GitHub repo: https://github.com/roborao/Swin-Brain-Age-Estimation.

We conduct experiments on different popular CNN models to compare performance with each model, such as VGG16 and ResNet18. We used ImageNet classification weights for our pretrained models, including VGG16, ResNet18 and Swin ViT models. The three networks are trained for 25 epochs with a batch size of 4. For training purposes, SGD optimizer was selected, with a learning rate of 0.001 for ResNet18. The convergence of training and validation loss are shown in Fig. 3.

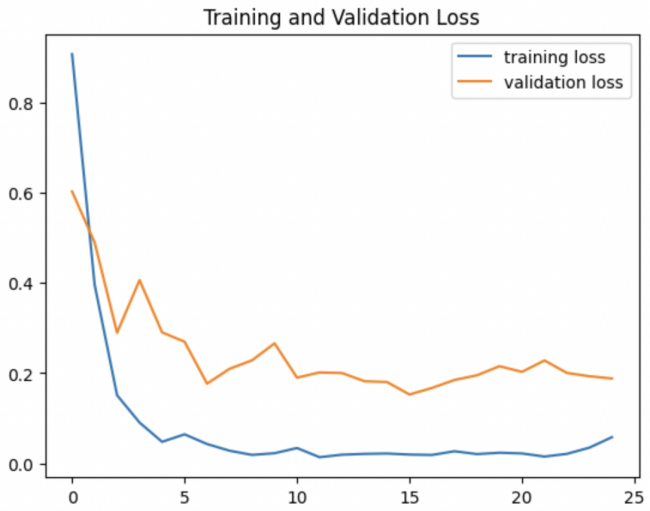

Fig. 3. ResNet18 Training and Validation Loss

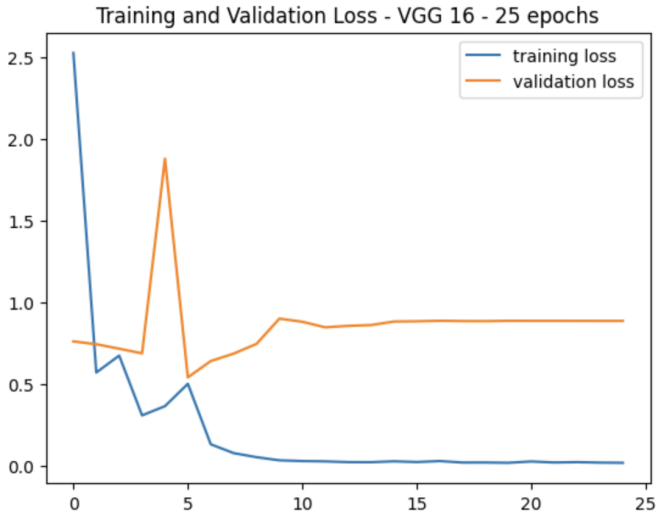

Fig. 4. VGG16 Training and Validation Loss

For the VGG16 model, we used Adam optimizer, with a learning rate of 0.0001. Its training and validation loss are shown in Fig. 4. Finally, for the Swin ViT model, we select a learning rate of 0.00001 with a learning rate decay of 0.05 every 5 epochs with the AdamW optimizer. We show the training and validation loss for the Swin ViT model below in Fig. 5 along with the training and validation accuracy in Fig. 6. Since this is a multi-class classification problem, cross entropy loss is used for the training optimization process in all models. The metric used to evaluate trained model is accuracy, which is the most common metric for classification problems. We also include F1-score, precision, and recall as other metrics.

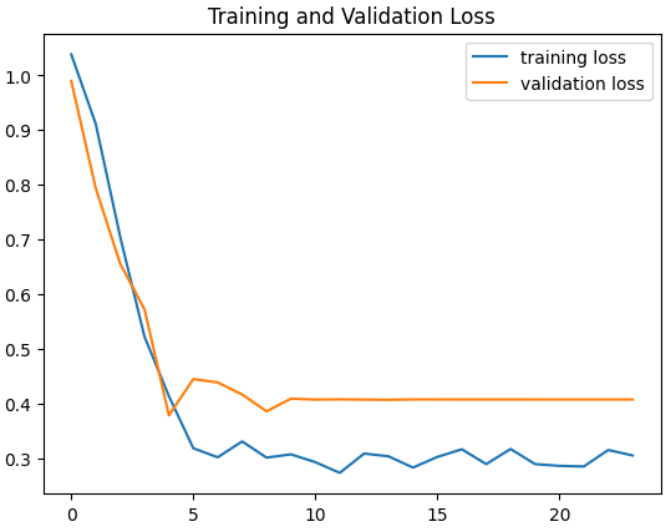

Fig. 5. Swin ViT Training and Validation Loss

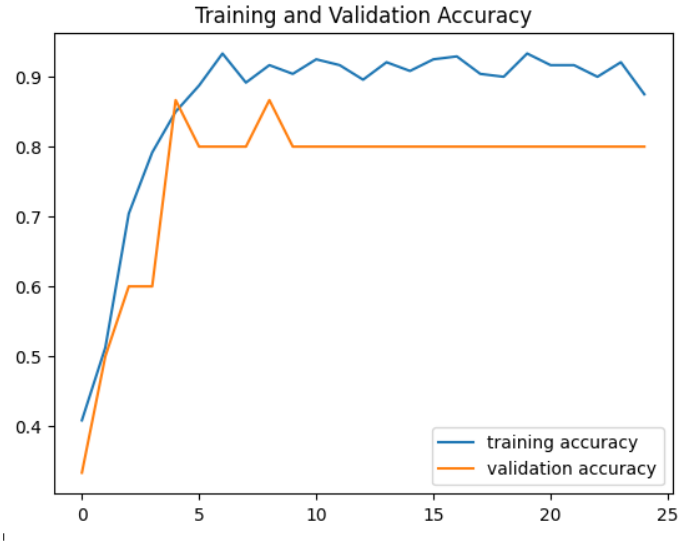

Fig. 6. Swin ViT Training and Validation Accuracy

After training, we achieved an 85% accuracy on test dataset when trained with ResNet18 architecture and an 82.5% accuracy when trained on VGG16 architecture. The training and validation losses were converging as expected during training process. When trained on Swin ViT, we achieved an accuracy of

92.5% on test dataset, a significant improvement in performance. The comparison of the test accuracies between the three tested models is shown below, in Table I.

Model	Classification Accuracy
VGG-16	0.825
ResNet-18	0.85
Swin ViT	**0.925**

TABLE I
CLASSIFICATION ACCURACY FOR THE TESTED MODELS

Table II shows the competitive values achieved by the Swin model for precision, recall, and f1-scores. Also presented is the Swin model's confusion matrix for the true and predicted labels in Fig. 7.

Label	Precision	Recall	F1-score
infant	0.8462	0.9167	0.8800
child	0.9412	0.8889	0.9143
adult	1.0000	1.0000	1.0000

TABLE II
CLASSIFICATION METRICS FOR SWIN MODEL

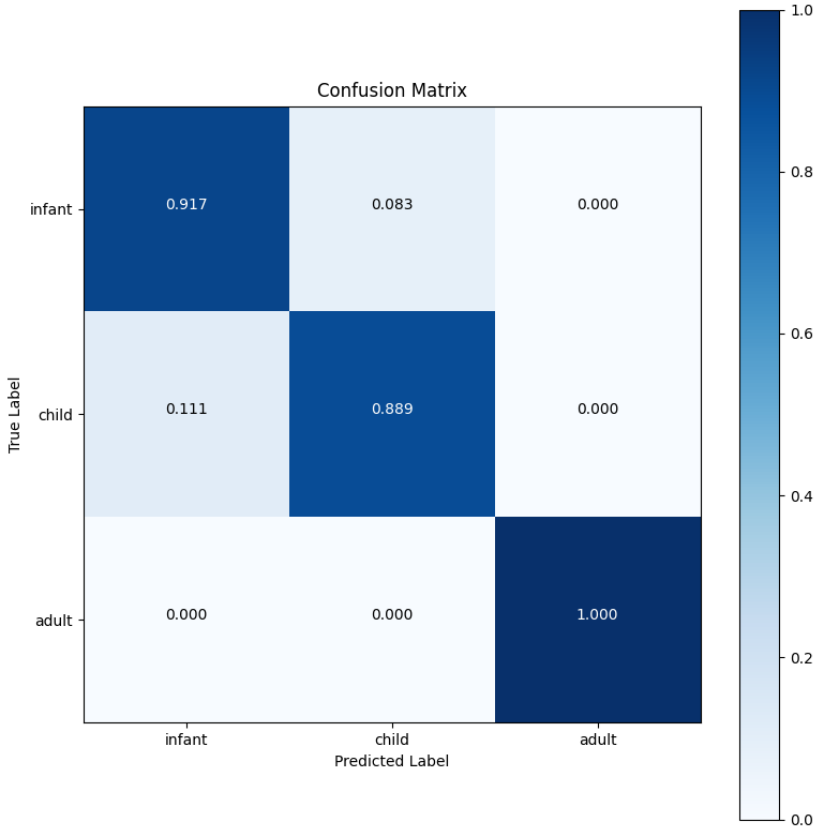

Fig. 7. Classification Confusion Matrix

Finally, we show qualitative samples of predictions against their ground truth in Fig. 8.

Fig. 8. Sample Brain Age Inference

V. DISCUSSION

We have clearly shown a significant improvement in classification accuracy by fine tuning the Swin ViT in our application of brain age estimation, achieving 92.5% test accuracy. While we hypothesized that ViT could offer noticeable improvement over CNN approaches, we must comment that it is still a surprising result because of the small dataset size. This result attests to the strength of the attention mechanism, even in imaging applications. Given additional data samples for a larger dataset size, we likely could have achieved much higher accuracy across the board for all three models. We would also like to note that additional experiments could be run to test the efficiency of these models. While the Swin ViT achieved greater performance than both of its CNN counterparts in our experiments, it is a generally known result that CNNs are still much more efficient than ViTs. In general, transformers are data-hungry and this gives rise to a practical issue to run into in real world applications: how can we maintain superior performance with less data? Of course, one possible solution is data augmentation, which we have exercised in our experiments, in order to gain ”additional” data for free. In summary, we have shown experimentally that our hypothesis that ViTs can be used to achieve better performance than CNN in brain age estimation holds true.

REFERENCES

[1] 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. 2020. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.* arXiv. doi: 10.48550/arXiv.2010.11929.

[2] Liu, Ze, et al. *Swin transformer: Hierarchical vision transformer using shifted windows.* Proceedings of the IEEE/CVF international conference on computer vision. 2021. arXiv. doi: 10.48550/arXiv.2103.14030

[3] Asiri AA, Shaf A, Ali T, Shakeel U, Irfan M, Mehdar KM, Halawani HT, Alghamdi AH, Alshamrani AFA, Alqhtani SM. *Exploring the Power of Deep Learning: Fine-Tuned Vision Transformer for Accurate and Efficient Brain Tumor Detection in MRI Scans.* Diagnostics (Basel). 2023 Jun 16;13(12):2094. doi: 10.3390/diagnostics13122094. PMID: 37370989; PMCID: PMC10297056.

[4] Wahlang, I., Maji, A. K., Saha, G., Chakrabarti, P., Jasinski, M., Leonowicz, Z., Jasinska, E. (2022). *Brain Magnetic Resonance Imaging Classification Using Deep Learning Architectures with Gender and Age.* Brain Sciences, 12(3), 391. doi: 10.3390/s22051766

[5] Dinsdale, N. K., Bluemke, E., Smith, S. M., Arya, Z., Vidaurre, D., Jenkinson, M., Namburete, A. I. (2020). *Learning patterns of the ageing brain in MRI using deep convolutional networks.* NeuroImage, 224, 117401. doi: 10.1016/j.neuroimage.2020.117401

[6] He, S., Gollub, R. L., Murphy, S. N., Perez, J. D., Prabhu, S., Pienaar, R., Robertson, R. L., Grant, P. E., Ou, Y. (2020). *Brain Age Estimation Using LSTM on Children's Brain MRI. In Medical Image Computing and Computer Assisted Intervention – MICCAI 2020.* Springer. doi: 10.1109/isbi45749.2020.9098356

[7] Illakiya, T., Karthik, R. (2023). *A Dimension Centric Proximate Attention Network and Swin Transformer for Age-Based Classification of Mild Cognitive Impairment From Brain MRI.* IEEE Access, 11, 120988-121002. doi: 10.1109/ACCESS.2023.3327821

[8] Deininger, L., Abbasi-Sureshjani, S., Stimpel, B., Schönenberger, S., Yuce, A., Ocampo, P., Korski, K., Gaire, F. (2022). *A comparative study between vision transformers and CNNs in digital pathology.* arXiv preprint arXiv:2206.00389. doi: doi.org/10.48550/arXiv.2206.00389

[9] Richardson, H., Lisandrelli, G, Riobueno-Naylor, A, Saxe, R. (2023). *MRI data of 3-12 year old children and adults during viewing of a short animated film.* OpenNeuro. doi: doi:10.18112/openneuro.ds000228.v1.1.1