Vision Transformers (ViT): Applying Vision Transformer to Brain Tumor MRI Classification

Ayobami Adebesin [1]

Abstract

In this report, we evaluate the effectiveness of Vision Transformers (ViT) as compared to traditional Convolutional Neural Networks (CNNs) for brain tumor classification in MRI images. We fine-tuned both a Vision Transformer model and a ResNet50 architecture on a brain tumor MRI dataset, comparing their performance across training time, inference speed, and classification accuracy metrics. This report investigates our central hypothesis: *Can a ViT outperform traditional CNNs on small medical datasets when properly fine-tuned?*

1. Introduction

The application of transformer architectures to computer vision represents a fundamental shift in how machines interpret visual data, particularly for critical healthcare applications like brain tumor classification in MRI scans. While convolutional neural networks (CNNs) are the standard, state-of-the-art architectures for vision problems and image classification, their localized receptive fields struggle to capture global contextual relationships-a limitation that becomes critical when analyzing complex anatomical structures in MRI scans where tumors may span dispersed regions. Vision Transformers (**?**) (ViTs) address this gap by leveraging self-attention mechanisms to model long-range dependencies across entire images, enabling holistic interpretation of spatial patterns essential for accurate tumor detection

1.1. Problem Statement and Motivation

The motivation for this study is primarily based on two challenges:

- Architectural limitations of CNNs: Traditional models like ResNet50 prioritize local feature extraction through convolutional filters, potentially missing subtle tumor signatures that require global contextual analysis.
- Clinical urgency for precision: Brain tumors account for 85-90% of primary central nervous system cancers, with MRI serving as the primary diagnostic tool. Even modest improvements in classification accuracy could significantly impact early intervention strategies and survival rates.

1.2. High level overview of the ViT architecture

Since the original transformer architecture is used for sequence modeling tasks, Vision transformer, using an encoder only architecture, reformulates this into a image classification task as follows:

- **Patch-based processing**: Images are divided into 16×16 pixel patches, converted into linear embeddings that preserver spatial relationships through positional encoding.
- **Positional Embeddings**: Spatial information is preserved through a learned positional embedding, as opposed to fixed or sinusoidal positional embedding commonly used in traditional transformers.
- **Multi-head self-attention**: A multi-head self-attention mechanism, similar to what was used in the original transformer architecture for sequence task, is used to learn complex spatial relationships and interactions between different regions of the input, enabling the model to capture both local and global contextual dependencies effectively. Each transformer layer dynamically weights interactions between all image patches, allowing the model to focus on diagnostically relevant regions regardless of their spatial proximity.

1.3. Clinical and Technical Significance

The transition to transformer-based models has some important implications as highlighted below:

- Diagnostic accuracy: Preliminary studies show ViTs achieving 98.13% accuracy on brain tumor classification, outperforming CNNs by 4-7% on some challenging datasets. This performance gap stems from ViTs' ability to correlate distal image features that often indicate tumor infiltration boundaries.

- Computational efficiency: Despite initial assumptions about complexity, optimized ViT implementations demonstrate comparable inference times to ResNet50 when processing high-resolution MRI scans.

- Generalization potential: Unlike CNNs that require extensive dataset-specific tuning, ViTs' pretraining on large natural image corpora allows effective transfer learning to medical domains with limited annotated data.

By addressing both the technical limitations of conventional architectures and the pressing clinical need for reliable diagnostic tools, this work positions transformer-based models as a transformative technology for medical imaging. The following sections detail how our implementation harnesses these architectural advantages while overcoming challenges specific to MRI analysis.

2. Methodology

In this section, we dive into the details of the vision transformer architecture and also review the details of the ResNet50 architecture.

2.1. Vision Transformer

Since transformers are originally designed to receive 1D sequence of token embeddings as input, the vision transformer adapts this formulation for a vision problem as follows: Let $x \in \mathbb{R}^{H\times W\times C}$ be a 2-D image with width W, height H and number of channels C. The image is divided into patches of size $x_p \in \mathbb{R}^{P\times P\times C}$, where P is the patch size. The total number of patches will be $N = \frac{H}{P} \times \frac{W}{P}$. Each image patch x_p is then flattened into a 1D vector $\mathbf{x_p} \in \mathbb{R}^{P^2C}$, $p = 1, 2, \ldots N$. These vectors are stacked together to form a matrix $X \in \mathbb{R}^{N\times P^2C}$ where each row represent the vectorized image patch. The matrix X is then projected into a D-dimensional embedding space via a trainable matrix $E \in \mathbb{R}^{P^2C\times D}$, given by $Z = XE$, to have the output $Z \in \mathbb{R}^{N\times D}$. The output is known as patch embeddings. A high-level overview is given in Fig 2.1

A special learnable embedding $z_{class} \in \mathbb{R}^D$ is prepended to the sequence, so that the matrix $Z \in \mathbb{R}^{(N+1)\times D}$. It acts like the [CLS] token in BERT and is used for classification. Because transformers are order-agnostic, we add a learnable positional embedding $E_{pos} \in \mathbb{R}^{(N+1)\times D}$ to Z so as to inject spatial information. So the final input to the transformer is a matrix of shape $(N + 1) \times D$.

This input $Z \in \mathbb{R}^{(N+1)\times D}$ is what will be passed into the encoder layer of the transformer, where the multi-head self-attention(MSA) mechanism will be used to compute attention scores for the input. The mechanism is the same

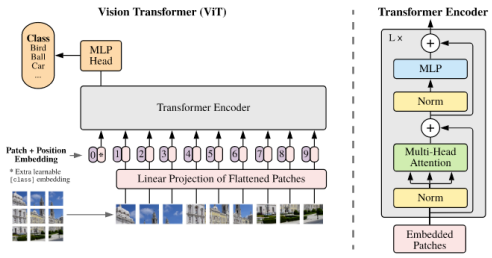

Figure 1. High-Level ViT Architecture

as described in the original Attention is All You Need paper. The output from this layer will be passed into the MLP layer which serves as the classification head with one hidden layer at pre-training and one single linear layer for fine-tuning.

Mathematically, the ViT architecture can be desribed as follows:

3. Dataset Description and Preprocessing

3.1. Dataset Overview

The study utilizes the Brain-Tumor-MRI-Dataset from Hugging Face. The dataset is freely available at https://huggingface.co/datasets/Hemg/Brain-Tumor-MRI-Dataset. The dataset comprises of $7,023$ axial MRI scans across four diagnostic classes:

- Glioma (23.1%)
- Meningioma (23.1%)
- Pituitary Tumor (25%)
- No Tumor (28.5%)

The near-balanced class distribution minimizes bias risks while preserving clinical relevance, reflecting real-world tumor incidence rates.

3.2. Key Challenges

The following are some of the challenges encountered with the dataset:

- Limited Data Scale: Despite being sizable for medical imaging, the dataset remains small compared to natural image corpora used for pretraining vision models.

- Subtle Pathological Features: Tumor boundaries often blend with healthy tissue, requiring models to detect submillimeter texture variations.

3.3. Preprocessing Pipeline

To enable fair model comparison, we utilized the following:

- Standardization:
  - Resized all images to 224×224 pixels
  - Converted grayscale images to RGB
  - Normalized pixels values using the mean and standard deviation
- Dataset Splitting:
  - Training: 6320 images (90%)
  - Validation: 703 images (10%

This preprocessing ensures compatibility with both Vision Transformer (ViT) and ResNet50 architectures while maintaining diagnostic information in downsampled images. The pipeline addresses MRI-specific challenges through resolution normalization and artifact suppression via controlled resizing techniques.

4. Implementation

The code for the implementation of the ViT fine-tuning process is attached with this report. The code consists of two IPython notebooks:

- `vit_finetune.ipynb`: This notebook contains the code that was used in fine-tuning the ViT model for our classification tasks. The pretrained model, **googlevit-base-patch15-224-in21k** was trained on the ImageNet dataset, and was fine-tuned by replacing the classifier with one that suits the need of our task.
- `resnet50_finetune.ipynb`: This notebook contains the code that was used in finetuning the ResNet50 model. The pretrained model, **microsoftresnet50** was also trained on the Imagenet dataset, and we finetuned it by replacing the classifier with one that suits our purpose.

4.1. Vision Transformer Configuration

Architecture: Patch Embeddings $\rightarrow$ Transformer Encoder $\rightarrow$ Classification head

Fine-Tuning Modifications:

- Added a classifier head that outputs 4 classes
- Input: 224×224 RGB patches (16×16 patch size)

Training Parameters:

- Optimizer: AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with weight decay 0.1
- Learning Rate: $2e^{-5}$
- Batch Size: 16 (limited by GPU memory constraints)
- Epochs: 10

4.2. ResNet50 Configuration

- Replaced final fully connected layer with 4-class output
- Global average pooling retained from original architecture
- Optimizer: AdamW, with weight decay 0.01
- Learning Rate: $3e^{-4}$
- Batch Size: 16 limited by GPU memory constraints)
- Epochs: 15

4.3. Computation Resources and Environment

This project was built using the Python programming language, HuggingFace transformers, and PyTorch framework in a controlled `conda` virtual environment to enable reproducibility. All computations use double precision. The model was built on a machine with 16-core GPU and utilizes PyTorch Metal Performance Shaders (MPS); a backend framework that enables accelerated training on Apple silicon GPUs.

5. Experimental Results

5.1. Quantitative Performance Comparison

Model	Val Acc	Training Time	Inference Time
ViT	98.5%	35 mins	0.02 secs
ResNet50	99.4%	31 mins	0.006 secs

Table 1. Performance comparison between Vision Transformer and ResNet50 on brain tumor MRI classification.

As shown in Fig 5.1, the ViT achieves a 98.5% validation accuracy while the ResNet50 achieves a 99.4% validation accuracy, indicating that the ResNet50 model performs higher. The results are not what was expected from our hypothesis, but it is possible that with careful fine-tuning, we can achieve a better accuracy with the ViT model. Our hypothesis suggests that the ViT model will outperform the traditional CNN model. However, in this experiment, this was not the outcome, but the performance seems to be on par with traditional CNNs, indicating that with more progress on the ViT architecture in the future, it might be possible that transformer-based models will outperform traditional CNNs. Fig 5.1 and Fig 5.1 shows the confusion matrix plot for both models.

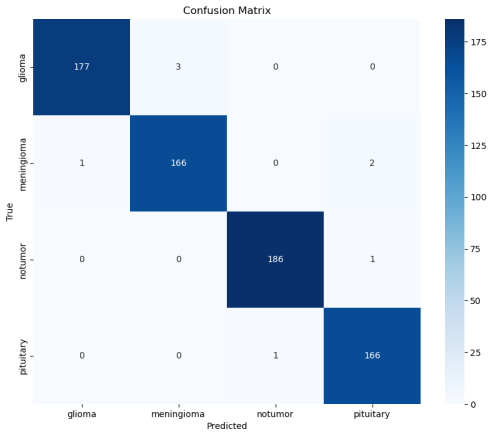

Figure 2. Confusion Matrix for ViT

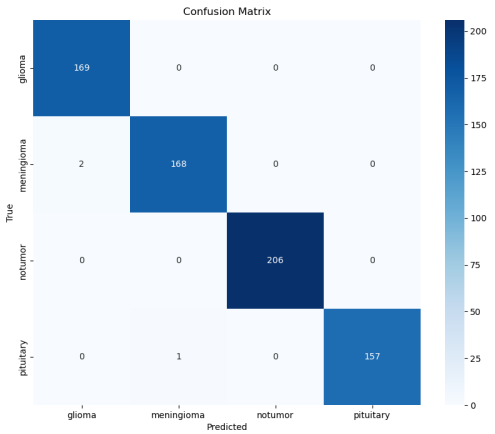

Figure 3. Confusion matrix for ResNet50