Brain Tumor Classification from MRI using Vision Transformers Ensembling

Sudhakar Tummala (sudhakar.t@srmap.edu.in)
SRM University AP https://orcid.org/0000-0001-5735-9418

Research Article

Keywords: brain tumor, MRI, vision transformer, diagnosis

Posted Date: April 26th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1593662/v1

Abstract

Automated classification of brain tumors plays an important role in supporting radiologists in decision making. Recently, Vision Transformer (ViT) based deep neural network architectures have gained attention in the computer vision research domain owing to the tremendous success of transformer models in natural language processing. However, studies involving vision transformers for various tasks in the medical imaging domain, including in the field of neuroimaging, are still growing. Many methods have been developed for the classification of brain tumors using traditional machine learning and deep learning methods. In particular, there are several convolutional neural network based transfer learning approaches for achieving good tumor classification accuracy. In this study, pretrained and finetuned ViT models on the ImageNet were adopted for the classification task. A brain tumor dataset from figshare consisting of 3064 T1-weighted contrast-enhanced (CE) magnetic resonance imaging (MRI) slices with meningioma, glioma, and pituitary tumor was used for cross-validation and testing of ensembled ViT models ability for 3-class classification task. The ensemble of all four ViT models B/16, B/32, L/16, and L/32, has demonstrated an overall testing accuracy of 98.7% at 384 × 384 resolution. Therefore, an ensemble of ViT models could be deployed for the computer-aided diagnosis of brain tumors based on T1w CE MRI leading to radiologist relief.

Introduction

Brain tumors (BT) are characterized by the abnormal growth of neural and glial cells in the brain. BT causes several medical conditions including loss of sensation, hearing and vision problems, headaches, nausea and seizures [1, 2]. There exist several types of brain tumors and the most prevalent cases include meningioma (originates from the membrane surrounding the brain) which is non-cancerous, glioma (starts from glial cells and spinal cord) and glioblastoma (grows from the brain) which are cancerous [3, 4]. Sometimes cancer can spread from other parts of the body which is called brain metastasis [5]. Pituitary tumor is another type of brain tumor that develops in the pituitary gland in the brain which primarily regulates other glands of the body [6]. Magnetic resonance imaging (MRI) is a versatile imaging method that enables to visualize inside the body noninvasively and it is in extensive use in the field of neuroimaging [7]. There exist several structural MRI protocols to visualize inside the brain but prime modalities include T1-weighted (T1w), T2-weighted, and T1w contrast-enhanced (CE) MRI. BTs appear with different pixel intensity contrasts in structural MRI images compared with neighboring normal tissues enabling clinical radiologists to diagnose the tumor [8].

There were several studies to classify brain tumors automatically using MRI images starting with traditional machine learning classifiers such as support vector machines (SVM), k-nearest-neighbor (kNN), and Random Forest from hand crafted features of the MRI slices [9–12]. With the rise of convolutional neural network (CNN) deep learning model architectures since 2012 and along with emerging advanced computational resources such as GPUs and TPUs, during the past decade, several methods have been proposed for the classification of brain tumors based on finetuning the existing state-of-the-art CNN models such as AlexNet, VGG16, ResNets, Inception, DenseNets, Xception, which were

already successful for various computer vision tasks [13–22]. These aforementioned pretrained CNN models based on localized convolutions demonstrated excellent performance in the brain tumor classification that were tested on different datasets [23–26]. CNNs generally have inductive bias i.e., the translation equivariance of the local receptive field, due to the inductive bias, the CNN models have issues learning long range information, and moreover, data augmentation is generally required for CNNs to improve their performance due to their dependency on local pixel variations during learning.

Lately, transformers [27] have become the de facto models for natural language processing. An adapted version of the transformer for images, the vision transformer (ViT), has been proposed in [28] and it seemingly performed superior to CNN models under a huge data regime as demonstrated by its improved performance when it was trained on JFT dataset with 300M images [28]. Therefore, to fully exploit the power of ViTs, a large amount of data is required and it may not be possible in medical imaging domains to collect such an amount of data. To deal with this, transfer learning approaches can be applied using pretrained and finetuned ViT models. These approaches were already successful in a few existing medical imaging diagnostics [29–33]. Hence, in this work, the ability of pretrained and finetuned ViT models both individually and in an ensemble manner is evaluated for the classification of meningioma, glioma and pituitary tumors from T1w CE MRI at 224 × 224 and 384 × 384 resolutions.

Related Work

ML and CNN based networks

Before the feasibility of using deep CNN based models in brain tumors classification, ML classifiers based on feature engineering from MRI images are the standard. In [9], several texture features extracted from MRI images were used to train SVM, kNN and extreme learning machine classifiers. In another study [11], tumor classification was conducted based on tumor shape, image intensity characteristics and rotation invariant texture features along with an SVM classifier and this method was applied to classify 102 types of brain tumors. In a study based on features extracted from structural MRI, diffusion-weighted MRI, and perfusion MRI, a four-class tumor classification system was developed using an SVM classifier [12].

With the rise of several state-of-the-art deep CNN models and the advent of transfer learning, neural network architectures have emerged as the standard for brain tumor classification from MRI. In [10], ensemble deep features extracted from 13 pretrained CNN models along with 9 machine learning classifiers are employed for improved classification. A Siamese network based tumor identification was performed based on GoogLeNet encodings and contrastive loss in a medical image retrieval study [13]. Similar GoogLeNet encodings along with ML classifiers were employed for brain tumor classification on the internet of medical things setup in other studies [22, 24]. Glioma classification using the data from the multimodal brain tumor image segmentation benchmark 2018 and the cancer imaging archive low grade glioma was performed using 2D mask regional CNN and 3D CNN models [15]. Since CNN models are also data hungry, variational autoencoders along with generative adversarial networks were used for synthetic

data generation and ResNet50 for tumor classification in a very recent study [18]. Using the figshare brain tumor dataset, transfer learning from VGG16, VGG19, ResNet50, and DenseNet21 models with four different optimization algorithms were implemented and the authors concluded that ResNet50 with Adadelta performed better among all [19]. Despite the success of CNN models, they have an inherent inductive bias which limits their performance towards unseen data when the object of interest in the image has different orientations and scales.

ViT based networks

ViT models proposed by [28] have a less inductive bias due to global patch-based learning and they learn more appropriate inductive biases specific to the requirement. Usage of ViT models for medical imaging diagnostics is sparse and still in its infancy because ViTs were recently introduced and they require large amounts of data and higher computational resources for training to exhibit their full potential. In [34], several pretrained and finetuned models on ImageNet21k and ImageNet2012 datasets with various patch sizes and the different number of multi-head self-attention layers allowing finetuning to a downstream task are provided and are openly available.

In [35], ViTs ability to classify breast cancer from ultrasound image is presented where the authors compared the performance of several pretrained and finetuned models and concluded that ViTs performed better than conventional CNNs; in particular, ViT-B/32 achieved superior performance among all. In another recent work [36], a ViT based explainable covid-19 and pneumonia classification model was developed from chest X-rays and computed tomography images. To address ViT demands in terms of large data and computational resources, a data efficient transformer was introduced based on regularization and data augmentation methods similar to CNNs [37]. Swin transformer is another variant of ViT based on shifted windows technique [38] and recently a Swin-Unet has been proposed for multi-organ and cardiac image segmentation tasks [39]. In another recent work [40], the Segtran model was developed for medical image segmentation tasks such as optic disc segmentation in digital fundus images, polyp segmentation in colonoscopy images and brain tumor segmentation using MRI based on squeeze-and-expansion transformers. More recent advances in ViT based models in various fields including the medical imaging field could be found here [41].

Methods

This section describes the dataset, the vision transformer architecture, computational infrastructure for model training, hyperparameter tuning using the validation set, and testing. The ViT models ensembling and the performance metrics employed are also discussed.

Dataset

An openly available dataset from figshare consists of 3064 T1-weighted CE MRI slices from 233 patients with meningioma or glioma or pituitary tumors. The images are available in all sagittal, coronal and axial directions with spatial resolutions of 512 × 512 or 256 × 256. More details about the dataset are available

at [42, 43]. A few MRI images from the figshare dataset are illustrated in Fig. 1. Further, a brief clinical description of the three types of tumors is given below.

Meningioma: these are mostly benign tumors originating from the arachnoid cap cells and often occur in older age individuals and females. These tumors account for 13–26% of all intracranial tumors [44].

Glioma: gliomas are the most frequent and primary intracranial tumors which are malignant. They represent 81% of all intracranial tumors which can cause significant mortality and morbidity [45].

Pituitary Tumor: it originates in the pituitary gland and is mostly benign. Since this gland regulates different hormones, tumors present in it may cause severe changes in the body. These tumors contribute to 10–15% of all intracranial tumors [3].

The number of images for each tumor category and the number of images used for training, validation and testing in a 70:10:20 ratio respectively are described in Table 1.

Table 1
Figshare dataset showing the number of MRI slices for each tumor category. MRI: magnetic resonance imaging, BT: brain tumor, N: number of images.

BT type	Total Images	Training	Validation	Testing
Meningioma	708	502	75	131
Glioma	1426	988	148	290
Pituitary Tumor	930	647	91	192
Total (N)	3064	2137	314	613

Vision transformer

The ViT proposed by [28] works by treating image patches as words to mimic the original transformer model developed for natural language processing tasks [27]. Although the original transformer model was the combination of both an encoder and a decoder, the ViT model has only an encoder in its architecture. In ViT, the input image I is $\mathrm{R}^{H\times W\times C}$, it is divided into N patches of size $P\times P\times C$ where $N=\frac{HW}{P^2}$ (*H*: height, *W*: width, *C*: number of channels). Afterward, linear embeddings are computed for these image patches and position embeddings are added to them to keep the patch positional information (Fig. 2). An extra learnable patch embedding is added for final classification by a multilayer perceptron (MLP) head. Further, these combined patch and position embeddings are fed to the transformer encoder model which has alternating layers of multiheaded self-attention and MLP blocks (Fig. 3). In this work, the pretrained and finetuned ViT base (B) and large (L) models: B/16, L/16, B/32 and L/32 (16 and 32 indicate square patch size) on ImageNet-21k and ImageNet-1k datasets respectively were used. Hence, the MRI images were resized to the resolution of 224 × 224 and 384 × 384. Since these

pretrained ViT models require three channels in the input and as the MRI slice has a single channel, the same grayscale MRI image is copied into the other two channels.

Like [class] in BERT [46], a learnable embedding is concatenated to the sequence of patch embeddings ($z_0^0 = I_{class}$). Mathematically, the working principle of ViT is given below in Equations (1)-(4). In Eq. (1), E_{pos} is the positional embeddings and $x_p^N E$ is the embedding of patch *N* which was a learnable linear projection. The first block of the transformer encoder layer starts with layer normalization (LN) followed by multi-head self-attention (MSA), and a residual connection follows that; the second block also starts with the LN layer followed by an MLP and a residual connection as shown in Fig. 3 and Equations (2) and (3). The MLP in the transformer block contains two fully connected layers with GELU (gaussian error linear unit) nonlinearity. The output of the final transformer encoder layer will be z_L^0 which is further layer normalized as described in Eq. (4) to get the final latent representation y (with dimension D) of the input image I. The MLP head or the final classification head is attached to this final latent representation (Fig. 2) during both pretraining and finetuning.

$$z_0 = \left[I_{class}; x_p^1 E; x_p^2 E; \ldots; x_p^N E\right] + E_{pos} E \in \mathrm{R}^{(P^2 . C) \times D}, E_{pos} \in \mathrm{R}^{(N+1) \times D} \quad (1)$$

$$z_l^{'} = MSA\left(LN\left(z_{l-1}\right)\right) + z_{l-1} l = 1 \ldots L \quad (2)$$

$$z_l = MLP\left(LN\left(z_l^{'}\right)\right) + z_l^{'} l = 1 \ldots L \quad (3)$$

$$y = LN\left(z_L^0\right) \quad (4)$$

More details about the pretraining and finetuning of ViT models on larger datasets are described in detail in [28].

Computational infrastructure

Google Colab Pro cloud environment which provides about 25 GB RAM along with nvidia T4 GPU accelerator was used. The model training, validation and testing were implemented in TensorFlow 2.8.0 which has *Keras* as a high-level API. The pretrained and finetuned ViT models available at the *vit-keras* module are used for the downstream task of 3-class classification of brain tumors from the figshare dataset. Custom Python scripts were written where and when necessary.

Model ensembling

To evaluate the ensemble models for class prediction, the procedure described in Equations (5) and (6) is followed. The softmax outputs of each model ($softmax_i$) are dot-wise added and finally divided by the number of individual models (*N*) to obtain the final output ($softmax_e$) of the ensemble classifier. Two

ensembling procedures are evaluated, where the first one is the ensemble of all models at 224 × 224 resolution and the second ensemble is combining all models at 384 × 384 resolutions.

$$softmax_e = \frac{1}{N}\sum_{i=1}^{N} softmax_i \quad (5)$$

$$finalclassprediction = argmax\left(softmax_e\right) \quad (6)$$

Performance metrics

Since it is a multi-class classification, sparse categorical cross-entropy was used as the loss metric and sparse categorical accuracy was used as the performance metric during training and validation. Confusion matrix and overall sparse categorical accuracy are used as model evaluation metrics during testing. The model's hyperparameters that were tuned are optimizer (*RMSprop/Adam/Adadelta*), learning rate (*lr*), number of epochs (*ne*), and mini-batch size (*mbs*). Optimization of the hyperparameters was conducted using the validation set. For calculating performance metrics on the test set, the hyperparameters that gave the best accuracy values during 5-fold cross-validation are considered.

Results

Initially, the image intensities were rescaled to have values between − 1 and 1, which is a requirement for ViT models. During training, all parameters of the ViT models were allowed to be finetuned. For the input image resolution of 224 × 224, the optimized hyperparameters with respect to validation accuracy are *Adam* optimizer with *lr* = 0.0001, *ne* = 25, and *mbs* = 16. B/16 model performed best at this resolution with a validation accuracy of 97.83%. For the rest of the models, performance at different hyperparameter combinations is given in Table 2 and the best hyperparameters and accuracy values are highlighted.

Similarly, at 384 × 384 resolution, the optimized hyperparameters for the best validation accuracy of 98.64% from L/16 model are *Adadelta* with *lr* = 0.1, *ne* = 10 and *mbs* = 8. *Adadelta* was solely the best optimizer at this resolution. The optimized hyperparameters and validation accuracies for all other models B/16, B/32, L/16, and L/32 are 98.10%, 98.04%, and 98.55% respectively. Due to computational constraints, training at 384 resolution is implemented with lower *mbs* values.

Table 2

Validation accuracy values for different optimizers and hyperparameters for ViT-B/16, ViT-B/32, ViT-L/16 and ViT-L/32 for both input image resolutions of 224 × 224 and 384 × 384. ViT: vision transformer, *ne* = number of epochs, *mbs* = mini-batch size, *lr* = learning rate. B: base, L: large.

Resolution	Optimizers & Hyperparameters	Validation accuracy in percentage			
		ViT-B/16	**ViT-B/32**	**ViT-L/16**	**ViT-L/32**
224 × 224	*RMSprop*				
	$\begin{cases} lr = 0.0001, ne = 25, mbs = 16 \\ lr = 0.0001, ne = 20, mbs = 32 \\ lr = 0.00005, ne = 15, mbs = 32 \end{cases}$	$\begin{cases} 96.20 \\ 96.41 \\ 97.06 \end{cases}$	$\begin{cases} 97.28 \\ 97.01 \\ 96.47 \end{cases}$	$\begin{cases} 96.10 \\ 96.47 \\ 95.92 \end{cases}$	$\begin{cases} 96.20 \\ 95.92 \\ 95.65 \end{cases}$
	Adam				
	$\begin{cases} lr = 0.0001, ne = 25, mbs = 16 \\ lr = 0.0001, ne = 20, mbs = 32 \\ lr = 0.00005, ne = 15, mbs = 32 \end{cases}$	$\begin{cases} 97.83 \\ 97.55 \\ 96.47 \end{cases}$	$\begin{cases} 95.92 \\ 96.74 \\ 96.74 \end{cases}$	$\begin{cases} 96.82 \\ 96.40 \\ 96.50 \end{cases}$	$\begin{cases} 97.25 \\ 96.20 \\ 97.25 \end{cases}$
	Adadelta				
	$\begin{cases} lr = 0.1, ne = 15, mbs = 16 \\ lr = 0.1, ne = 20, mbs = 32 \\ lr = 0.05, ne = 15, mbs = 32 \end{cases}$	$\begin{cases} 97.25 \\ 97.01 \\ 97.55 \end{cases}$	$\begin{cases} 96.01 \\ 96.01 \\ 96.20 \end{cases}$	$\begin{cases} 97.28 \\ 97.25 \\ 97.55 \end{cases}$	$\begin{cases} 97.28 \\ 97.25 \\ 96.20 \end{cases}$
384 × 384	*RMSprop*				
	$\begin{cases} lr = 0.0001, ne = 15, mbs = 8 \\ lr = 0.0001, ne = 10, mbs = 16 \\ lr = 0.00005, ne = 10, mbs = 8 \end{cases}$	$\begin{cases} 97.31 \\ 96.60 \\ 97.63 \end{cases}$	$\begin{cases} 97.55 \\ 97.21 \\ 96.74 \end{cases}$	$\begin{cases} 97.40 \\ 96.95 \\ 97.60 \end{cases}$	$\begin{cases} 96.51 \\ 96.60 \\ 97.60 \end{cases}$
	Adam				
	$\begin{cases} lr = 0.0001, ne = 15, mbs = 8 \\ lr = 0.0001, ne = 10, mbs = 16 \\ lr = 0.00005, ne = 10, mbs = 8 \end{cases}$	$\begin{cases} 97.30 \\ 97.54 \\ 96.90 \end{cases}$	$\begin{cases} 97.11 \\ 96.65 \\ 97.01 \end{cases}$	$\begin{cases} 96.82 \\ 97.40 \\ 97.70 \end{cases}$	$\begin{cases} 97.01 \\ 97.40 \\ 96.60 \end{cases}$
	Adadelta				
	$\begin{cases} lr = 0.1, ne = 10, mbs = 8 \\ lr = 0.1, ne = 15, mbs = 16 \\ lr = 0.05, ne = 10, mbs = 8 \end{cases}$	$\begin{cases} 97.10 \\ 97.80 \\ 98.10 \end{cases}$	$\begin{cases} 98.04 \\ 97.83 \\ 96.84 \end{cases}$	$\begin{cases} 97.90 \\ 97.50 \\ 98.64 \end{cases}$	$\begin{cases} 98.55 \\ 97.60 \\ 98.01 \end{cases}$

The test accuracy values for both the input image resolutions of 224 × 224 and 384 × 384 for all ViT models are given in Table 3. ViT-B/16 performed well among all with an overall accuracy of 97.06% at 224 × 224. Similarly, at the resolution of 384 × 384, the ViT-L/32 emerged as the single best classifier with overall test accuracy of 98.21%.

Table 3
Test accuracy values are given in percentages for ViT-B/16, ViT-B/32, ViT-L/16 and ViT-L/32 for both resolutions of 224 × 224 and 384 × 384. ViT: vision transformer, B: base, L: large.

Resolution	Test accuracy in percentage			
	ViT-B/16	ViT-B/32	ViT-L/16	ViT-L/32
224 × 224	97.06	96.25	96.74	96.01
384 × 384	97.72	97.87	97.55	98.21

The performance of the average ensembling on the test set is given in Table 4. The ensembling of the models at 224 × 224 resolution resulted in an overall accuracy of 97.71% and the overall test accuracy of the ensemble model at 384 × 384 resolution is 98.7%.

Table 4
Test accuracy are values given in percentages for ensemble classification at a) resolution of 224 × 224, b) resolution of 384 × 384. ViT: vision transformer.

Ensembled Models	Test accuracy in percentage
All ViT models at 224 × 224 resolution	97.71
All ViT models at 384 × 384 resolution	98.70

The performance of ViT models on the test set in the form of confusion matrices is given in Figs. 4 and 5 for 224 × 224 and 384 × 384 resolutions respectively. The number of false predictions is higher for meningioma and glioma compared to the pituitary tumor. A similar trend was observed at the two resolutions. However, the number of false predictions is relatively lower at 384 × 384 resolution. Figure 6 shows the confusion matrices for the ensemble models performance at both resolutions on the test set. The number of false predictions for the ensemble model at 384 resolution was just eight and the ensemble model achieved 100% accuracy in the identification of glioma.

Discussion

In this study, the ability of pretrained and finetuned ViT models is investigated both individually and in an ensemble manner for 3-class classification of brain tumors including meningioma, glioma and pituitary tumor from T1w CE MRI. In general, all ViT models demonstrated the ability to classify with validation and test accuracies above 97% during most scenarios (refer to Tables 2 and 3). Based on the hyperparameter tuning using the validation set, the performance of all the models is good irrespective of

the choice of the model hyperparameters, such as optimizer, *lr*, *ne*, and *mbs* which indicates that the ViT models are robust across different hyperparameter settings; however, the *Adadelta* optimizer outperformed all other optimizers at 384 × 384 resolution. Nevertheless, to evaluate the performance of the models on the test set, the models that yielded the highest accuracy values on the validation set was considered which is the standard procedure. The individual model's performance on both the validation and test sets is slightly better at the image resolution of 384 × 384 compared to 224 × 224, which could be because the general performance of the ViT models is better at higher resolutions, as indicated by the experiments from [28]. Similarly, the ensemble model's performance at 384 × 384 was better than that of the ensemble model's performance at 224 × 224 because average ensembling was used and the ensemble model's performance depends on the individual model's performance in the group.

Comparing the performance of the ViT models with previous studies based on the same dataset given in Table 5, the ensemble of ViTs at 384 × 384 resolution performed better, with an overall test accuracy of 98.7%. Based on the confusion matrices on the test set from all the models at both input image resolutions (Figs. 4 and 5), meningioma has a higher number of misclassifications than glioma and pituitary tumors possibly because there could be feature overlapping between the image encodings of meningioma and glioma, as well as meningioma and pituitary tumor. Previous studies have documented a similar trend in misclassification in test set results [19, 22]. Our study outperformed all previous studies based on custom CNNs and transfer learning methods indicating that the pretrained and finetuned ViT models are superior to CNN based models. The only study that performed marginally better was from [19]; however, the number of false predictions in [19] was 9 whereas, in our study, the number of false predictions was 8 using ensemble model with 384 × 384 resolution, as shown in Fig. 6B.

Table 5
Previous related work using **figshare** dataset and performance comparison in terms of overall test accuracy. ViT: vision transformer.

Work	Method	Image resolution	Training	Test Accuracy
J Cheng [42]	GLCM-BoW	512 × 512	80%	91.28%
MR Ismael [47]	DWT-2D Gabor	512 × 512	70%	91.90%
A Pashaei [48]	CNN-ELM	512 × 512	70%	93.68%
P Afshar [49]	CapsuleNet	128 × 128	-	90.89%
S Deepak [22]	CNN-SVM-kNN	224 × 224	80%	97.80%
O Polat [19]	Transfer Learning	224 × 224	70%	99.02%
B Ahmad [18]	GAN-VAEs	512 × 512	60%	96.25%
Our method	Ensemble of ViTs	224 × 224	70%	97.71%
		384 ×384	70%	98.70%

During training, all the model parameters starting from the patch embeddings layer were allowed to be finetuned since based on a few experiments conducted by freezing the initial layers including some transformer encoder block layers of the ViT models, the validation and test accuracies are around a couple of percentage points lower than the accuracy values obtained by unfreezing parameters of all layers. Even though the model's performance improved at 384 × 384 resolution, training at this resolution was computationally demanding and hence implemented in a TPU environment. Further, the performance of the ViTs at the original input image resolution of 512 × 512 may become better and this hypothesis could be investigated in a high-level computing environment. Furthermore, the cross-validated models from this study can be finetuned to deal with other brain tumor datasets. In addition, in a future study, it could be interesting to investigate the ability of other vision transformer variants such as swin vision transformers [38], data-efficient vision transformers [37], and transformer in transformer models [50] for the brain tumor classification from MRI. A python notebook with specific code and cross-validated ViT models pertaining to this study can be provided upon reasonable request.

Conclusions

The performance of the ensemble model at 384 × 384 resolution is on par and better than previous CNN models for the classification of brain tumors from MRI achieving an overall test accuracy of 98.7%. Using the same ensemble model, the test classification accuracy for gliomas is 100%. Therefore, computer-aided diagnosis of brain tumors from T1w CE MRI using the ensemble of finetuned ViT models can be an alternative to manual diagnosis by a clinical radiologist.

Declarations

Competing Interests:

The authors have no competing interests to declare.

Funding:

No funding was received for conducting this study.

Compliance with Ethical Standards

This research study was conducted retrospectively using human subject data made available in open access by Figshare. Ethical approval was not required as confirmed by the license attached with the open access data.

References

1. S. Rasheed, K. Rehman, M.S.H. Akash, An insight into the risk factors of brain tumors and their therapeutic interventions, Biomed. Pharmacother. 143 (2021).

https://doi.org/10.1016/J.BIOPHA.2021.112119.

2. I. Sánchez Fernández, T. Loddenkemper, Seizures caused by brain tumors in children, Seizure. 44 (2017) 98–107. https://doi.org/10.1016/J.SEIZURE.2016.11.028.
3. M. Chintagumpala, A. Gajjar, Brain tumors, Pediatr. Clin. North Am. 62 (2015) 167–178. https://doi.org/10.1016/J.PCL.2014.09.011.
4. K. Herholz, K.J. Langen, C. Schiepers, J.M. Mountz, Brain tumors, Semin. Nucl. Med. 42 (2012) 356–370. https://doi.org/10.1053/J.SEMNUCLMED.2012.06.001.
5. A. Boire, P.K. Brastianos, L. Garzia, M. Valiente, Brain metastasis, Nat. Rev. Cancer. 20 (2020) 4–11. https://doi.org/10.1038/S41568-019-0220-Y.
6. G. Kontogeorgos, Classification and pathology of pituitary tumors, Endocrine. 28 (2005) 27–35. https://doi.org/10.1385/ENDO:28:1:027.
7. M. Viallon, V. Cuvinciuc, B. Delattre, L. Merlini, I. Barnaure-Nachbar, S. Toso-Patel, M. Becker, K.O. Lovblad, S. Haller, State-of-the-art MRI techniques in neuroradiology: principles, pitfalls, and clinical applications, Neuroradiology. 57 (2015) 441–467. https://doi.org/10.1007/S00234-015-1500-1.
8. J.E. Villanueva-Meyer, M.C. Mabray, S. Cha, Current Clinical Brain Tumor Imaging, Neurosurgery. 81 (2017) 397–415. https://doi.org/10.1093/NEUROS/NYX103.
9. K. Kavin Kumar, T. Meera Devi, S. Maheswaran, An Efficient Method for Brain Tumor Detection Using Texture Features and SVM Classifier in MR Images, Asian Pac. J. Cancer Prev. 19 (2018) 2789–2794. https://doi.org/10.22034/APJCP.2018.19.10.2789.
10. J. Kang, Z. Ullah, J. Gwak, MRI-Based Brain Tumor Classification Using Ensemble of Deep Features and Machine Learning Classifiers, Sensors (Basel). 21 (2021) 1–21. https://doi.org/10.3390/S21062222.
11. E.I. Zacharaki, S. Wang, S. Chawla, D.S. Yoo, R. Wolf, E.R. Melhem, C. Davatzikos, Classification of brain tumor type and grade using MRI texture and shape in a machine learning scheme, Magn. Reson. Med. 62 (2009) 1609. https://doi.org/10.1002/MRM.22147.
12. S. Shrot, M. Salhov, N. Dvorski, E. Konen, A. Averbuch, C. Hoffmann, Application of MR morphologic, diffusion tensor, and perfusion imaging in the classification of brain tumors using machine learning scheme, Neuroradiology. 61 (2019) 757–765. https://doi.org/10.1007/S00234-019-02195-Z.
13. S. Deepak, P.M. Ameer, Retrieval of brain MRI with tumor using contrastive loss based similarity on GoogLeNet encodings, Comput. Biol. Med. 125 (2020) 103993. https://doi.org/10.1016/J.COMPBIOMED.2020.103993.
14. Z.N.K. Swati, Q. Zhao, M. Kabir, F. Ali, Z. Ali, S. Ahmed, J. Lu, Brain tumor classification for MR images using transfer learning and fine-tuning, Comput. Med. Imaging Graph. 75 (2019) 34–46. https://doi.org/10.1016/J.COMPMEDIMAG.2019.05.001.
15. Y. Zhuge, H. Ning, P. Mathen, J.Y. Cheng, A. V. Krauze, K. Camphausen, R.W. Miller, Automated glioma grading on conventional MRI images using deep convolutional neural networks, Med. Phys. 47 (2020) 3044–3053. https://doi.org/10.1002/MP.14168.

16. R. Pomponio, G. Erus, M. Habes, J. Doshi, D. Srinivasan, E. Mamourian, V. Bashyam, I.M. Nasrallah, T.D. Satterthwaite, Y. Fan, L.J. Launer, C.L. Masters, P. Maruff, C. Zhuo, S.C. Johnson, J. Fripp, N. Koutsouleris, D.H. Wolf, R. Gur, R. Gur, J. Morris, M.S. Albert, H.J. Grabe, S.M. Resnick, R. Nick Bryan, D.A. Wolk, R.T. Shinohara, H. Shou, C. Davatzikos, Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan, (2019). https://doi.org/10.1016/j.neuroimage.2019.116450.
17. M.A. Naser, M.J. Deen, Brain tumor segmentation and grading of lower-grade glioma using deep learning in MRI images, Comput. Biol. Med. 121 (2020). https://doi.org/10.1016/J.COMPBIOMED.2020.103758.
18. B. Ahmad, J. Sun, Q. You, V. Palade, Z. Mao, Brain Tumor Classification Using a Combination of Variational Autoencoders and Generative Adversarial Networks, Biomedicines. 10 (2022). https://doi.org/10.3390/BIOMEDICINES10020223.
19. Ö. Polat, C. Güngen, Classification of brain tumors from MR images using deep transfer learning, J. Supercomput. 2021 777. 77 (2021) 7236–7252. https://doi.org/10.1007/S11227-020-03572-9.
20. H.A. Khan, W. Jue, M. Mushtaq, M.U. Mushtaq, H.A. Khan, W. Jue, M. Mushtaq, M.U. Mushtaq, Brain tumor classification in MRI image using convolutional neural network, Math. Biosci. Eng. 2020 56203. 17 (2020) 6203–6216. https://doi.org/10.3934/MBE.2020328.
21. M.M. Badža, M.C. Barjaktarović, Classification of Brain Tumors from MRI Images Using a Convolutional Neural Network, Appl. Sci. 2020, Vol. 10, Page 1999. 10 (2020) 1999. https://doi.org/10.3390/APP10061999.
22. S. Deepak, P.M. Ameer, Brain tumor classification using deep CNN features via transfer learning, Comput. Biol. Med. 111 (2019) 103345. https://doi.org/10.1016/J.COMPBIOMED.2019.103345.
23. E.U. Haq, H. Jianjun, K. Li, H.U. Haq, T. Zhang, An MRI-based deep learning approach for efficient classification of brain tumors, J. Ambient Intell. Humaniz. Comput. 2021. (2021) 1–22. https://doi.org/10.1007/S12652-021-03535-9.
24. A. Sekhar, S. Biswas, R. Hazra, A.K. Sunaniya, A. Mukherjee, L. Yang, Brain tumor classification using fine-tuned GoogLeNet features and machine learning algorithms: IoMT enabled CAD system, IEEE J. Biomed. Heal. Informatics. PP (2021). https://doi.org/10.1109/JBHI.2021.3100758.
25. N.S. Shaik, T.K. Cherukuri, Multi-level attention network: application to brain tumor classification, Signal, Image Video Process. 2021. (2021) 1–8. https://doi.org/10.1007/S11760-021-02022-0.
26. M.F. Alanazi, M.U. Ali, S.J. Hussain, A. Zafar, M. Mohatram, M. Irfan, R. Alruwaili, M. Alruwaili, N.H. Ali, A.M. Albarrak, Brain Tumor/Mass Classification Framework Using Magnetic-Resonance-Imaging-Based Isolated and Developed Transfer Deep-Learning Model, Sensors (Basel). 22 (2022). https://doi.org/10.3390/S22010372.
27. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention Is All You Need, Adv. Neural Inf. Process. Syst. 2017-December (2017) 5999–6009. https://doi.org/10.48550/arxiv.1706.03762.

28. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, (2020). https://doi.org/10.48550/arxiv.2010.11929.
29. Y. Wu, S. Qi, Y. Sun, S. Xia, Y. Yao, W. Qian, A vision transformer for emphysema classification using CT images, Phys. Med. Biol. 66 (2021). https://doi.org/10.1088/1361-6560/AC3DC8.
30. B. Gheflati, H. Rivaz, Vision Transformer for Classification of Breast Ultrasound Images, (2021). https://doi.org/10.48550/arxiv.2110.14731.
31. F. Shamshad, S. Khan, S.W. Zamir, M.H. Khan, M. Hayat, F.S. Khan, H. Fu, Transformers in Medical Imaging: A Survey, (2022). https://doi.org/10.48550/arxiv.2201.09873.
32. J. Wang, Z. Fang, N. Lang, H. Yuan, M.Y. Su, P. Baldi, A multi-resolution approach for spinal metastasis detection using deep Siamese neural networks, Comput. Biol. Med. 84 (2017) 137–146. https://doi.org/10.1016/J.COMPBIOMED.2017.03.024.
33. Y. Dai, Y. Gao, F. Liu, TransMed: Transformers Advance Multi-modal Medical Image Classification, Diagnostics. 11 (2021). https://doi.org/10.48550/arxiv.2103.05940.
34. A. Steiner, A. Kolesnikov, X. Zhai, R. Wightman, J. Uszkoreit, L. Beyer, How to train your ViT? Data, Augmentation, and Regularization in Vision Transformers, (n.d.). https://github.com/rwightman/pytorch-image-models. (accessed March 10, 2022).
35. B. Gheflati, H. Rivaz, VISION TRANSFORMERS FOR CLASSIFICATION OF BREAST ULTRASOUND IMAGES, (n.d.).
36. A.K. Mondal, A. Bhattacharjee, P. Singla, A.P. Prathosh, xViTCOS: Explainable Vision Transformer Based COVID-19 Screening Using Radiography, IEEE J. Transl. Eng. Heal. Med. 10 (2022). https://doi.org/10.1109/JTEHM.2021.3134096.
37. H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, F. Ai, Training data-efficient image transformers & distillation through attention, (2020). https://doi.org/10.48550/arxiv.2012.12877.
38. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, (2021). https://doi.org/10.48550/arxiv.2103.14030.
39. H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, M. Wang, Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation, (2021). http://arxiv.org/abs/2105.05537 (accessed March 10, 2022).
40. S. Li, X. Sui, X. Luo, X. Xu, Y. Liu, R. Goh, Medical Image Segmentation Using Squeeze-and-Expansion Transformers, (2021). https://github.com/askerlee/segtran. (accessed March 10, 2022).
41. K. Islam, Recent Advances in Vision Transformer: A Survey and Outlook of Recent Work, (2022). https://doi.org/10.48550/arxiv.2203.01536.
42. J. Cheng, W. Huang, S. Cao, R. Yang, W. Yang, Z. Yun, Z. Wang, Q. Feng, Enhanced Performance of Brain Tumor Classification via Tumor Region Augmentation and Partition, PLoS One. 10 (2015). https://doi.org/10.1371/JOURNAL.PONE.0140381.

43. J. Cheng, W. Yang, M. Huang, W. Huang, J. Jiang, Y. Zhou, R. Yang, J. Zhao, Y. Feng, Q. Feng, W. Chen, Retrieval of Brain Tumors by Adaptive Spatial Pooling and Fisher Vector Representation, PLoS One. 11 (2016). https://doi.org/10.1371/JOURNAL.PONE.0157112.
44. C. Marosi, M. Hassler, K. Roessler, M. Reni, M. Sant, E. Mazza, C. Vecht, Meningioma, Crit. Rev. Oncol. Hematol. 67 (2008) 153–171. https://doi.org/10.1016/J.CRITREVONC.2008.01.010.
45. Q.T. Ostrom, H. Gittleman, L. Stetson, S.M. Virk, J.S. Barnholtz-Sloan, Epidemiology of gliomas, Cancer Treat. Res. 163 (2015) 1–14. https://doi.org/10.1007/978-3-319-12048-5_1.
46. J. Devlin, M.W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, NAACL HLT 2019–2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf. 1 (2018) 4171–4186. https://doi.org/10.48550/arxiv.1810.04805.
47. M.R. Ismael, I. Abdel-Qader, Brain Tumor Classification via Statistical Features and Back-Propagation Neural Network, IEEE Int. Conf. Electro Inf. Technol. 2018-May (2018) 252–257. https://doi.org/10.1109/EIT.2018.8500308.
48. A. Pashaei, H. Sajedi, N. Jazayeri, Brain tumor classification via convolutional neural network and extreme learning machines, 2018 8th Int. Conf. Comput. Knowl. Eng. ICCKE 2018. (2018) 314–319. https://doi.org/10.1109/ICCKE.2018.8566571.
49. P. Afshar, K.N. Plataniotis, A. Mohammadi, Capsule Networks for Brain Tumor Classification Based on MRI Images and Coarse Tumor Boundaries, ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc. 2019-May (2019) 1368–1372. https://doi.org/10.1109/ICASSP.2019.8683759.
50. K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, Y. Wang, Transformer in Transformer, (2021). https://doi.org/10.48550/arxiv.2103.00112.

Figures

Figure 1

MRI images from the figshare dataset are shown in sagittal, coronal and axial cut planes for meningioma, glioma and pituitary tumors.

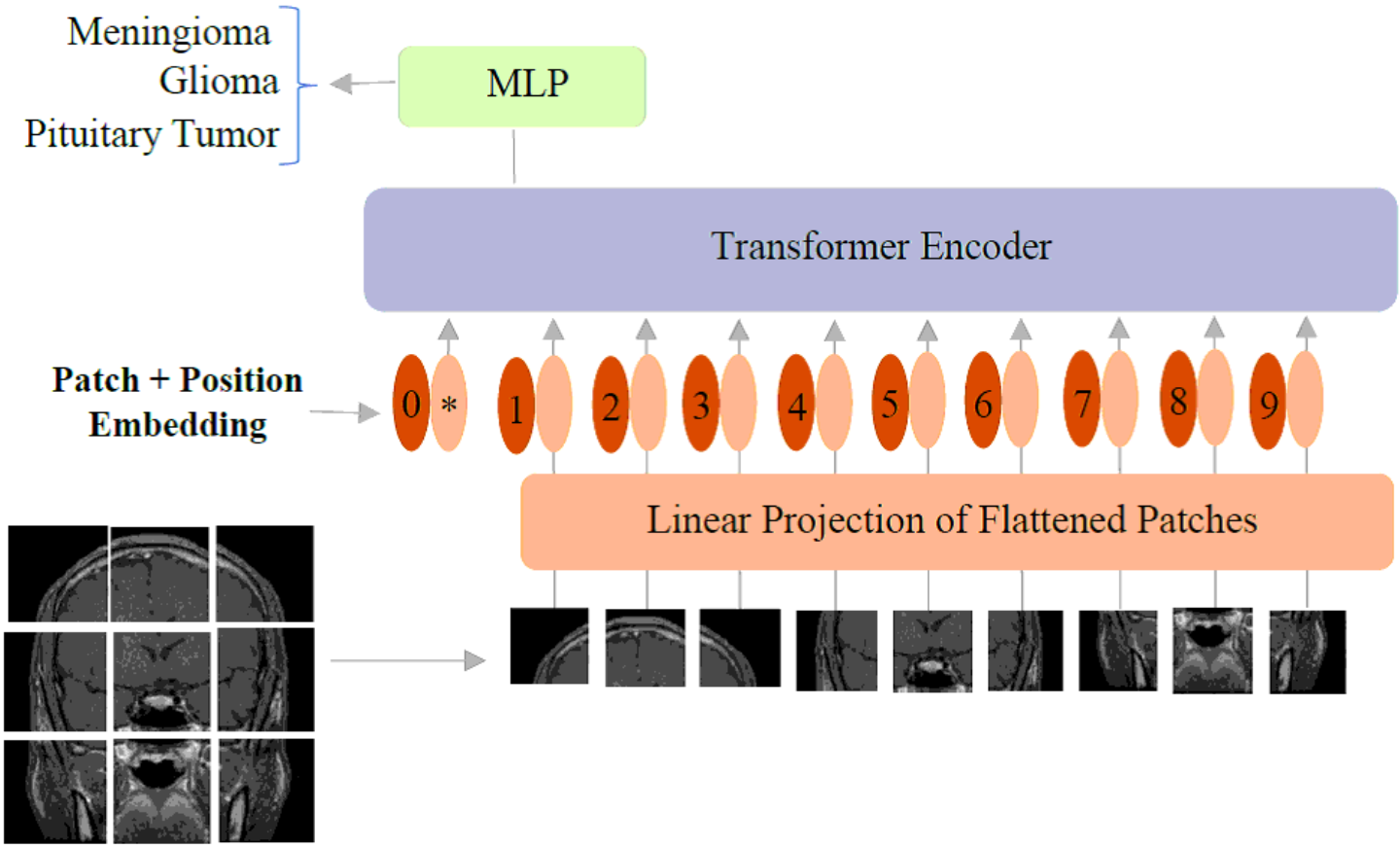

Figure 2

Vision transformer model adopted for classification of brain tumors from MRI. MLP: multilayer perceptron. *is extra learnable patch embedding to be used by the final classification head.

Figure 3

The vision transformer encoder with multi-head self-attention. LN: layer normalization, MLP: multilayer perceptron, L_x: transformer encoder '*x*' at layer L.

Figure 4

Confusion matrix for classification of three types of tumors on the test set using ViT models A) B/16, B) B/32, C) L/16, and D) L/32 at the image resolution of 224 × 224.

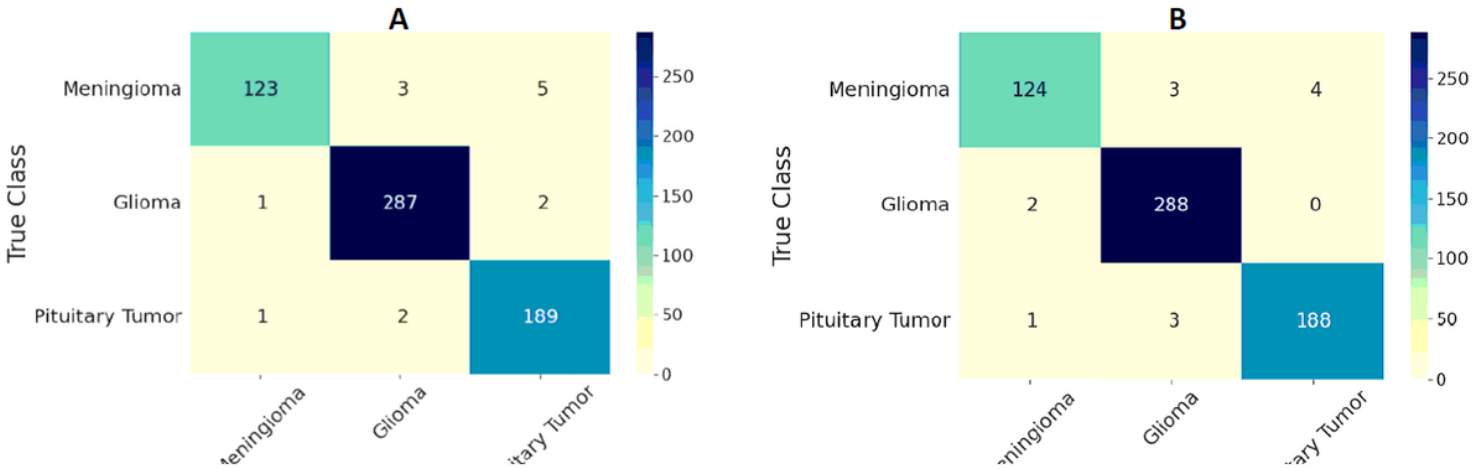

Figure 5

Confusion matrix for classification of three types of tumors on the test set using ViT models A) B/16, B) B/32, C) L/16, and D) L/32 at the image resolution of 384 × 384.

Figure 6

Confusion matrix for classification of three types of tumors on the test set using an ensemble of ViT models B/16, B/32, L/16, and L/32 at a) 224 × 224 resolution, b) 384 × 384 resolution.