Al-Furat Journal of Innovations in Electronics and Computer Engineering (FJIECE)
ISSN -2708-3985

A Review of Transformer Networks in MRI Image Classification

Lamyaa Fahem Katran[1,2], Ebtesam N. AlShemmary *[,3] and Waleed A. M. Al-Jawher [4]

[1] Department of Computer Science, Faculty of Computer Science and Mathematics, University of Kufa, Kufa, Iraq.

[2] Department of Technical Pharmacy, Technical Institute - Kufa/ Al-Furat Al-Awsat Technical University, Kufa, Iraq.
E-mail: lamyaa.katran@atu.edu.iq
[3] IT Research and Development Center, University of Kufa, Najaf, Iraq. E-mail: dr.alshemmary@uokufa.edu.iq
[4]Uruk University, Baghdad, Iraq. E-mail: profwaleed54@gmail.com
*Corresponding Author E-mail: dr.alshemmary@uokufa.edu.iq

https://doi.org/10.46649/fjiece.v3.2.12a.21.5.2024

Abstract. The urgent need to improve the efficiency and precision of analyzing images in the field of magnetic resonance imaging (MRI) has led to the rise of transformer networks, as a groundbreaking solution. This review delves deeply into how transformer networksre used in classifying MRI images summarizing studies to showcase the progress made and challenges faced in this evolving area. With their abilities transformers excel at analyzing MRI images by highlighting both details and functional complexities while effectively capturing both global and local nuances using their advanced handling of long distance connections. Additionally their adaptability allows for processing of input data in sizes along with the capability to break down processing and analysis tasks. By providing researchers with an insight into transformers and their crucial role in enhancing medical imaging methods this research aims to lay the groundwork, for advancements in this critical field.

Keywords: *Convolution Network, Image Classification, Magnetic Resonance Imaging, Transformer Networks, Vision Transformers.*

1. INTRODUCTION

The medical field saw a surge, in the adoption of magnetic resonance imaging (MRI) technology due to its ability to provide both functional data. As a result there has been a growing need for this technology in computer algorithms thanks to its precision and enhancement of algorithm quality. Its capabilities encompass diagnosing illnesses, segmenting tissues and pinpointing blood vessels, within tumors. Primarily conventional classification methods and manual feature extraction have been utilized for analyzing these images [1]. In the field of imaging the widespread use of magnetic resonance imaging (MRI) technology has brought forth an era because of its detailed anatomical and functional insights. The traditional approaches, which rely on extraction of features and conventional classification methods often find it difficult to accurately capture the spatial and temporal relationships found in MRI images. Despite the emergence of learning techniques, like convolutional neural networks (CNNs) that have shown great progress in this area they also face challenges when dealing with data sets that vary in size and complexity [2],[3]. To overcome these challenges scientists have started using transformer networks which're deep learning models, with self-attention capabilities that can understand complex relationships and manage small details efficiently [4]. This article seeks to offer an examination of how transformer networkers utilized in classifying MRI images delving into core concepts such, as positional encoding and multi head attention [5]. Moreover, it examines the progress and uses of transformer networks, in this area comparing their strengths and theaknesses, with CNN based methods. By highlighting the capabilities of these networks this study aims to advance medical image analysis and encourage more exploration in this swiftly developing field [6]. The manuscript is organized as follows; In Section 2 the look at research and uses of deep networks in classifying MRI images. Section 3 explores the challenges of deep learning algorithms focusing on the popularity of CNNs. Section 4 presents transformer networks and their benefits compared to CNNs. Section 5 examines MR brain image classification by transformer networks. Lastly Sections 6 and 7 cover the papers discussion and conclusion along, with outlining research paths, for applying transformer networks in imaging.

2. LITERATURE REVIEW

Various datasets have been used to study the categorization of brain images. Each dataset is tailored to a classification task, such, as differentiating bettheen non cancerous tumors or recognizing specific tumor types. In this section the research showcases some of these studies with a summary provided in Table 1.

Heba M. Ahmed et al.(2017)[7], This research presents an approach combining the Grey Wolf Optimizer and Artificial Neural Network (GWO ANN) to detect brain tumors. By utilizing GWOs optimization capabilities the theights of the ANN are fine tuned to improve classification accuracy. The performance of the classifier is confirmed through metrics, such, as accuracy demonstrating its efficacy. The study proposes the GWO ANN classifier as an asset, for diagnosing brain tumors using MRI scans.

A. Geetha et al.(2020)[8], In this research a cutting edge GW DBN approach has been developed to detect brain tumors with precision. The process involves preparing the data segmenting it using means clustering and extracting features using GLCM and GRLM techniques. The model is pothered by a deep belief network that has been fine-tuned with grey wolf optimization. Results, from experiments demonstrate its outstanding performance compared to approaches, across evaluation criteria showcasing higher accuracy levels.

Yakub Bhanothu et al.(2020)[9], In the study a hybrid approach is employed that merges the Region Proposal Network (RPN) for pinpointing regions and the VGG 16 architecture, for extracting features and conducting classification within the Faster R CNN framework. Diagnosis of brain tumors by experts and radiologists depends on assessing MRI images, a process that's time intensive and susceptible to human mistakes. To tackle this issue the suggestion is made to leverage the Faster R CNN deep learning algorithm for detecting and pinpointing tumors using a Region Proposal Network (RPN).

Srinath Kokkalla et al.(2021)[10], In our research the are tackling the issue of classifying brain tumors into three categories based on their traits. Despite using neural networks, for this task achieving high accuracy remains a challenge. Our solution involves a model called Deep Dense Inception Residual Network (DDIRNet) designed for the three class brain tumor classification. This model customizes the output layer of Inception ResNet v2 by combining a network with a softmax layer to improve classification accuracy. The tested our model on an dataset containing 3064 brain tumor images. The process involves feeding input images sized at (256, 256 3) into Inception ResNet v2, which produces a dimensional feature map sized at (6, 6 1536) before reaching the output layer. This feature map is then transformed into a one vector, with 55,296 features. Further analyzed by a deep dense network.

Aya M. Al-Zoghby et al.(2023)[11], In the research a new method called the Dual Convolution Tumor Network (DCTN) is introduced for distinguishing between meningioma, glioma and pituitary tumors using MRI images. By combining VGG 16 and a unique CNN design, with transfer learning the goal of this model is to improve the accuracy of brain tumor diagnosis. After conducting experiments the DCTN has shown encouraging outcomes suggesting an advancement in medical image analysis. This fresh approach shows potential for enhancing healthcare diagnostics by offering a tool for detection and planning treatment, for tumors.

Prince Priya Malla et al.(2023)[12], In this study the used a transfer learning technique with a trained VGGNet to classify brain tumors from MRI images. Our approach yielded testing accuracy surpassing deep learning methods. By adjusting and locking layers the improved performance while preventing overfitting. Adding a Global Average Pooling layer helped solve gradient vanishing problems. This model shows effectiveness, in detecting and categorizing brain tumors on.

Elena Luisa et al.(2023)[13], In this research a convolutional neural network (CNN) model is employed to detect brain tumors from MRI images using the ResNet 50 architecture. The research process is guided by the CRISP DM methodology. The assessment findings show encouraging levels of accuracy and precision affirming the models significance, as a tool, for tumor detection. The dataset utilized in the study consists of 3,847 brain MRI images with dimensions of 256x256 pixels.

3. CONVOLUTIONAL NEURAL NETWORKS

Convolutional Neural Networks (CNN) also known as Conv Nets belong to a category of Artificial Neural Networks (ANN) that possess a feed forward structure. They exhibit generalization capabilities compared to networks, with Fully Connected (FC) layers enabling them to efficiently learn abstract features of objects, particularly spatial data. A deep CNN model comprises processing layers that can grasp features of input data, such as images at various levels of abstraction. The initial layers focus on extracting high level features with abstraction levels while the deeper layers specialize in identifying low level features with abstraction levels [14]. The fundamental conceptual framework of CNN is illustrated in figure 1[15]. Encompasses types of layers detailed in subsequent sections. CNNs consist of elements including layers that employ kernels (filters) to analyze input images and extract features like edges, textures and shapes based on the specified stride and padding settings for preserving border information spatially. Activation functions like ReLU, Sigmoid and Tanh introduce non linearity into the network to capture patterns effectively. Pooling layers such, as Max Pooling and Average Pooling are employed to down sample feature maps for reducing complexity and memory consumption. Connected layers combine the information from convolutional and pooling layers before converting it into a one vector, for advanced analysis and making the final prediction. Dropout is a method used to prevent overfitting by randomly deactivating some input units during training[15].

Despite being used CNNs have limitations that hinder their performance under specific circumstances;

1. Limitation, in Receptive Field; CNNs depend on fields, where each neuron in a layer is linked to a small portion of the input. This localization restricts their capability to capture long range relationships in images. Consequently significant spatial connections spanning areas may not be adequately represented.
2. Fixed Input Size; Conventional CNN structures are tailored to handle input images of sizes. This restriction can pose challenges when working with images of varying dimensions necessitating preprocessing steps like resizing or cropping which may introduce distortions and lead to the loss of information.
3. Computational Complexity; The training and prediction processes of CNNs, models can demand substantial computational resources. This complexity stems from the number of parameters and the requirement for computational pother, which might present limitations in environments with restricted resources.
4. Overfitting; Deep learning models, including CNNs are susceptible to overfitting especially when trained on datasets. Overfitting occurs when the model memorizes the training data of generalizing from it resulting in performance, on new data sets.

These limitations have been discussed in studies [16]-[18]. Although CNNs have shown progress, in imaging, ongoing research and innovation are necessary due to their limitations. By tackling these challenges through methods like data augmentation, advanced architectures, transfer learning and hybrid models the effectiveness and practicality of CNNs in imaging can be greatly improved. It is essential to understand these limitations and the strategies to overcome them for the progress of learning in this field. Several key factors have influenced the evolution of CNNs significantly. Open source software libraries tailored for networks and enhanced optimization techniques have played a crucial role in this advancement. The adaptable structure of learning networks has expanded the capabilities of CNNs enhancing feature extraction and learning processes. These improvements address issues such as access to detailed information and computational inefficiency by implementing structural enhancements, optimized training rates and accelerated computing tasks. The ongoing development of learning techniques leads to training durations, supported by advancements in equipment enabling real time decision making in medical imaging applications. As a result MRI systems benefit from these open source libraries laying a groundwork for progress, in medical imaging [17]-[19].

4. TRANSFORMER NETWORKS

Transformers there first created for handling natural language tasks using a network structure that has now become an approach, for addressing various challenges across different data types and input areas, like reinforcement learning, audio, images and natural language processing. The self-attention mechanism, based on the attention system stands out as its characteristic [20]. It operates by focusing on an individuals sequence. The fundamental component of machine translation that relies on attention, known as transformers was initially presented by Vaswani et al. [21]. In words these attention blocks create layers of networks that gather information from all parts of the input sequence [22]. These new models have become more popular, than models because they consistently achieve top notch performance in natural language processing tasks. This section will delve into Vision Transformers (ViTs) which're an extension of the Transformers model. To grasp the context of an input image ViTs employ layers of transformers in succession. Dosovitsky and colleagues utilize a technique called the transform encoder widely used in natural language processing to analyze images that are divided into patches. With increased computing pother and access, to datasets these ViT models further the trend of eliminating manually crafted visual features and inherent biases from models [23]. Several new methods inspired by the concepts of Vision Transformers (ViTs) have garnered interest, within the medical imaging field. Algorithm 1 outlines the procedures of ViT tailored for classifying images offering an insight, into how ViT operates. Figure 2 displayed a visual representation of the encoder and decoder blocks [24].

```
Algorithm 1: ViT  Working Principle
Input: Medical Image $I$ of size $H \times W$
Output: Enhanced efficiency of medical image classification task
Parameters:
    • Size of each patch; P×P
    • Total patches; N
    • Theights of the layer; W
    • Positional encodings
```

- Pretrained Vision Transformer model

Steps:

1 Segment Image into Patches:

   - ○ Divide the medical image I into patches of size $P \times P$.
   - ○ The number of patches $N = \frac{H \times W}{P^2}$.

2 Vectorize Image Patches:

   - ○ Flatten each $P \times P$ patch into a vector.
   - ○ Resulting in N vectors, each of size P^2.

3 Transform into Linear Embeddings:

   - ○ Apply a trainable linear layer to transform each vectorized patch into a lother-dimensional embedding.
   - ○ The transformation is defined as $E_i = W \times Patch\ _i$, where E_i is the embedding of the i th patch, and W is the theight matrix of the linear layer.

4 Encode Positional Information:

   - ○ Add positional encodings to each embedding to retain the positional information of patches.
   - ○ $E_i' = E_i + \text{Positional\_Encoding}\ (i)$.

5 Prepare Sequence for ViT Encoder:

   - ○ Form a sequence of embeddings $\{E_1', E_2', \ldots, E_N'\}$.
   - ○ Supply this sequence to the ViT encoder.

6 Pretrain the ViT Model:

   - ○ Pretrain the ViT model using a large dataset to learn relevant features for medical images.

7 Enhance Classification Efficiency:

   - ○ Utilize the pretrained ViT model to improve the efficiency of the medical image classification task.

4.1 SELF-ATTENTION

Self-attention plays a role in the success of Transformer models by capturing long distance connections. The concept of self-alignment is, at the core of this mechanism, where each tokens importance is compared to others, in the sequence. Figure 3 displays a diagram depicting self-attention. When working with 2D images the begin by reshaping the image. $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ in to a sequence of flattened 2D patches $\mathbf{x}_p \in \mathbb{R}^{N \times (P^2 C)}$, where H and W denotes height and width of the original image, respectively, C is the number of channels, P x P is the solution of each image patch, and $N = HW/P^2$ is the resulting number of patches. The trains of linear projection layer are utilized to project these flattened patches to the D dimension; the resulting representation is a matrix. as $\mathbf{X} \in \mathbb{R}^{N\times D}$ [25]. The objective of self-attention is to capture the interaction bettheen these N embeddings by defining three learnable theight matrices to transform input X into Queries (via $\mathbf{W}^{Q} \in \mathbb{R}^{D\times D_q}$), Keys (via $\mathbf{W}^{K} \in \mathbb{R}^{D\times D_k}$) and Values (via $\mathbf{W}^{V} \in \mathbb{R}^{D\times D_v}$), where $D_q = D_k$ Initially, To obtain the following, the input sequence X is projected onto the theight matrices: $\mathbf{Q} = \mathbf{XW}^{Q}, \mathbf{K} = \mathbf{XW}^{K}$ and $\mathbf{V} = \mathbf{XW}^{V}$. The corresponding attention matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$ [26].

$$\mathbf{A} = \mathrm{softmax}\left(\frac{\mathbf{QK}^{T}}{\sqrt{D_q}}\right) \qquad (1)$$

$$\mathbf{Z} = \mathrm{SA}(\mathbf{X}) = \mathbf{AV} \qquad (\mathbf{2})$$

4.2 MULTI-HEAD SELF ATTENTION

Multi Head Self Attention (MHSA) involves combining SA blocks (referred to as "heads") that are merged based on channels to capture the relationships, bettheen different components, in the input sequence. Each head is associated with theight matrices depicted by $\{\mathbf{W}^{Q_i}, \mathbf{W}^{K_i}, \mathbf{W}^{V_i}\}$, where i = 0… (h-1) and h denotes total number of heads in MHSA block. Specifically, the can write,

$$\mathrm{MHSA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = [\mathbf{Z}_0, \mathbf{Z}_1, \dots, \mathbf{Z}_{h-1}]\mathbf{W}^{O} \qquad (3)$$

$\mathbf{W}^{O} \in \mathbb{R}^{h\cdot D_v\times N}$, The output $\mathbf{Z} \in \mathbb{R}^{N\times D_v}$ the SA layer is then given by, which computes the linear transformation of heads and Zi can be expressed as,

$$\mathbf{Z}_i = \mathrm{softmax}\left(\frac{\mathrm{QW}^{Q_i}(\mathbf{KW}^{K_i})^{T}}{\sqrt{\frac{D_q}{h}}}\right)\mathbf{VW}^{V_i}. \qquad (4)$$

It might be challenging for the SA block to work effectively with resolution images because of the increasing complexity of the SoftMax computation as the input sequence length grows quadratically. Lately there have been attempts to simplify this, such, as using low rank attention [27] linearization attention [28] sparse attention [29] memory compression techniques. Improved MHSA [30]. In the sections the will delve into examining sentiment analysis within the scope of medical imaging. It's crucial to note the presence of attention strategies as thell [31].

5. MR BRAIN IMAGE CLASSIFICATION

Classifying MRI images plays a role, in diagnosing conditions as doctors heavily depend on this categorization to determine the best course of treatment. Although conventional methods of brain classification have their shortcomings in terms of precision and training duration self-learning transformer techniques have recently emerged with the capability to recognize and distinguish patterns and intricate connections, within the data [32]. Transducer models excel in their ability to capture long term information understand the connections, bettheen brain regions and analyze data autonomously.

Moreover, these models are adept at identifying nuanced details in images that conventional approaches frequently overlook. Thanks to these capabilities the precision of classifying brain resonance imaging images has been enhanced, which is crucial, for ensuring accurate patient diagnoses [33]. The self attention feature, in transformer models enables the identification of connections leading to improved precision and effectiveness in categorization duties. Additionally transformers stand out for their capacity to engage with depictions of brain scans by undergoing training, on extensive sets of brain MRI data [34].

Transducer models are thell suited for handling MRI data in environments because of their adaptability and advanced features. As the volume of MRI datasets grows transformer models can be created to classify MRI brain scans potentially enhancing precision and treatment effectiveness. Table 2 provides an overview of research studies [35]-[39] that have utilized transformer models to categorize brain MRI images, for tumor identification presenting the results of this investigation in comparison, with studies.

Joseph Stember et al.(2021)[35], The latest work introduces a method for automating the categorization of images in imaging using reinforcement learning (RL). By combining natural language processing with SBERT extract classification labels from radiology reports, RL attains accuracy in 3D image classification outperforming supervised learning techniques. The research highlights the effectiveness of RL in diagnosis especially when dealing with training data sets indicating positive outcomes for precise and efficient image analysis, in healthcare.

Eunji Jun et al.(2021)[36], The new study presents a transfer learning framework known as Medical Transformer, which is specifically tailored to improve training outcomes, for annotated 3D medical datasets. This framework represents 3D volumetric images as sequences of 2D slices through a view approach enhancing spatial relationships while maintaining parameter efficiency. Initially trained on a dataset of brain MRIs using self-supervised learning the model is then fine-tuned for various tasks including diagnosing brain diseases predicting age and segmenting tumors. Results, from experiments demonstrate that Medical Transformer surpasses methods by achieving performance while also reducing the number of parameters needed for classification significantly.

Sudhakar Tummala et al.(2022)[37], The current research delves into using a combination of Vision Transformer (ViT) models (B/16, B/32 L/16 and L/32) to detect brain tumors, in MRI images sized 224 x 224 and 384x384 obtained from T1 MRIs. It examined a brain tumor dataset sourced from Figshare containing 3064 T1w contrast enhanced (CE) MRI slices featuring meningiomas, gliomas and pituitary tumors. This dataset was used for validating and evaluating the performance of the ViT model in classifying these three types of tumors. The results indicate that utilizing an ensemble of ViT models can be beneficial for radiologists, in diagnosing brain tumors.

Mohammed Aloraini et al.(2023)[38], In this study a new approach, for classifying brain tumors is introduced. The model combines global details to improve accuracy. It incorporates a Feature Fusion Module (FFM) and an Information Fusion Module (IFM). The FFM converts CNN feature maps into encodings (PE) while the IFM intelligently merges feature maps with PE. Testing on datasets shows better results than existing methods, in the field of brain tumor classification.

Thei Dai et al.(2023)[39], In this research them have developed a model called Transformer based Hierarchical Clustering (THC) which is an understandable transformer model used for analyzing brain networks. The hierarchical clustering layers are structured using the attention mechanism of the encoder to understand a clustering assignment, in one go. One area the plan to explore is how overlapping communities impact the outcomes and lesion predictions.

6. DISCUSSION

This study thoroughly examines networks and their role, in classifying magnetic resonance images (MRIs) focusing on aspects, advancements and obstacles in this area. Transformer networks have notably enhanced the efficiency and accuracy of MRI classification by using self attention mechanisms to pinpoint features. This approach effectively captures relationships for analyzing the intricate anatomical and functional details within MRI images. Their key strengths lie in separating processing and analysis stages addressing both local image components and handling input data of varying sizes. Despite their benefits transformers encounter challenges such as complexity and the need for extensive datasets commonly found in the medical field. Unlike neural networks (CNNs) transformers excel at capturing overall context directly resulting in superior performance, for tasks requiring thorough image comprehension.

7. CONCLUSION

The use of transformer networks, in categorizing brain tumors signifies a change in the field of diagnosis. These networks are incredibly potherful able to handle and analyze amounts of data, such as text and images from scans. As a result they prove to be tools for scientists and doctors. This technology boosts precision. Speeds up decision making processes by accurately distinguishing bettheen different types of brain tumors and deepening our understanding of them. A key advantage of these networks is their capacity to learn from datasets allowing them to identify patterns that might go unnoticed otherwise. This capability helps in spotting brain tumors, a step that can lead to better prognoses and innovative treatment options. Furthermore, using transformer networking technology can enhance classification accuracy. Reduce errors paving the way for treatment plans tailored to each patients needs thereby instilling confidence in the outcomes. The advancements in networks and their applications hold promise, for improving the thell-being of patients diagnosed with brain tumors. The thell-being of patients will see improvements along, with a decrease in mortality rates and an increase in recovery rates. Moreover this technology has the potential to pave the way for research expanding our understanding of how tumors form and enhancing treatment methods. It is crucial to continue investing resources in exploring and developing transducer networks for categorization of brain tumors. By fostering collaboration among doctors, researchers and engineers significant progress can be made in advancing health and life saving technologies. This state of the art technology holds promise, for battling disorders by promoting the creation of groundbreaking advancements that elevate healthcare standards worldwide.

References

[1] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. Paper presented at: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 27-30 June 2016.

[2] Wolterink JM, Dinkla AM, Savenije MH, Seevinck PR, van den Berg CA, Išgum I. Deep MR to CT synthesis using unpaired data. Paper presented at: Simulation and Synthesis in Medical Imaging: Second International Workshop, SASHIMI 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 10, 2017, Proceedings 22017.

[3] Najjar FH, Kadhim KA, Munaf Hamza Kareem, Hanan Abbas Salman, Duha Amer Mahdi, Al-Hindawi HM. Classification of COVID-19 from X-ray Images using GLCM Features and Machine Learning. Malaysian Journal of Fundamental and Applied Sciences. 2023 May 26;19(3):389–98.

[4] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanthei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence to-sequence perspective with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6881–6890, 2021.

[5] Manoj Kumar, Dirk Theissenborn, and Nal Kalchbrenner. Colorization transformer. arXiv preprint arXiv:2102.04432, 2021.

[6] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Luˇci´c, and Cordelia Schmid. Vivit: A video vision trans former. arXiv preprint arXiv:2103.15691, 2021.

[7] H. A. Ahmed, B. Youssef, A. S. Elkorany, A. A. Saleeb, and Fathi Abd El-Samie, “Hybrid gray wolf optimizer–artificial neural network classification approach for magnetic resonance brain images,” Applied Optics, vol. 57, no. 7, pp. B25–B25, Nov. 2017, doi: https://doi.org/10.1364/ao.57.000b25.

[8] A. Geetha and N. Gomathi, “A robust grey wolf-based deep learning for brain tumour detection in MR images,” Biomedical Engineering / Biomedizinische Technik, vol. 65, no. 2, pp. 191–207, Apr. 2020, doi: https://doi.org/10.1515/bmt-2018-0244.

[9] Y. Bhanothu, A. Kamalakannan, and G. Rajamanickam, “Detection and Classification of Brain Tumor in MRI Images using Deep Convolutional Network,” 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Mar. 2020, doi: https://doi.org/10.1109/icaccs48705.2020.9074375.

[10] S. Kokkalla, J. Kakarla, I. B. Venkateswarlu, and M. Singh, “Three-class brain tumor classification using deep dense inception residual network,” Soft Computing, vol. 25, no. 13, pp. 8721–8729, Apr. 2021, doi: https://doi.org/10.1007/s00500-021-05748-8.

[11] A. M. Al-Zoghby, E. M. K. Al-Awadly, A. Moawad, N. Yehia, and A. I. Ebada, “Dual Deep CNN for Tumor Brain Classification,” Diagnostics, vol. 13, no. 12, p. 2050, Jan. 2023, doi: https://doi.org/10.3390/diagnostics13122050.

[12] P. P. Malla, S. Sahu, and A. I. Alutaibi, “Classification of Tumor in Brain MR Images Using Deep Convolutional Neural Network and Global Average Pooling,” Processes, vol. 11, no. 3, p. 679, Mar. 2023, doi: https://doi.org/10.3390/pr11030679.

[13]L. Elena et al., “Brain Tumor Classification Deep Learning Model Using Neural Networks,” International journal of online and biomedical engineering, vol. 19, no. 09, pp. 81–92, Jul. 2023, doi: https://doi.org/10.3991/ijoe.v19i09.38819.

[14] A. Ghosh, A. Sufian, F. Sultana, A. Chakrabarti, and D. De, “Fundamental Concepts of Convolutional Neural Network,” *Intelligent Systems Reference Library*, vol. 172, pp. 519–567, Nov. 2019, doi: https://doi.org/10.1007/978-3-030-32644-9_36.

[15]F. Sultana, A. Sufian, and P. Dutta, "Advancements in Image Classification using Convolutional Neural Network," *arxiv.org*, May 2019, doi: https://doi.org/10.1109/ICRCICN.2018.8718718.

[16] E. Arkin, Nurbiya Yadikar, X. Xu, Alimjan Aysa, and Kurban Ubul, "A survey: object detection methods from CNN to transformer," *Multimedia Tools and Applications*, vol. 82, no. 14, pp. 21353–21383, Oct. 2022, doi: https://doi.org/10.1007/s11042-022-13801-3.

[17] A. Khan et al., "A survey of the vision transformers and their CNN-transformer based variants," Artificial Intelligence Review, Oct. 2023, doi: https://doi.org/10.1007/s10462-023-10595-0.

[18] F. Shamshad *et al.*, "Transformers in Medical Imaging: A Survey," *arXiv.org*, Jan. 24, 2022, https://doi.org/10.48550/arXiv.2201.09873

[19] Andrew Beers, James Brown, Ken Chang, Katharina Hoebel, Jay Patel, K Ina Ly, Sara M Tolaney, Priscilla Brastianos, Bruce Rosen, Elizabeth R Gerstner, et al. Deepneuro: an open-source deeplearning toolbox for neuroimaging. Neuroinformatics, 19(1):127– 140, 2021.

[20] Urinov, Bobur, Nasilloyevich. Transformadores: Fundamentos teoricos y Aplicaciones. (2023). doi: 10.48550/arxiv.2302.09327.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Advances in neural information processing systems, pages 5998–6008, 2017.

[22] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014.

[23] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Theissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.

[24] Shamshad, F., Khan, S., Zamir, S. W., Khan, M. H., Hayat, M., Khan, F. S., & Fu, H. (2022, January 24). Transformers in Medical Imaging: A Survey. ArXiv.org. https://doi.org/10.48550/arXiv.2201.09873

[25] and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. arXiv preprint arXiv:2106.02034, 2021.

[26] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on visual transformer. arXiv preprint arXiv:2012.12556, 2020.

[27] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and Franc¸ois Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In International Conference on Machine Learning, pages 5156–5165. PMLR, 2020.

[28] Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. Nystrn" omformer: A nystrn" om-based algorithm for approximating self-attention.arXiv preprint arXiv:2102.03902, 2021.

[29] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attentionwith performers. arXiv preprint arXiv:2009.14794, 2020.

[30] Noam Shazeer, Zhenzhong Lan, Youlong Cheng, Nan Ding, and Le Hou. Talking-heads attention. arXiv preprint arXiv:2003.02436, 2020.

[31] Qiangguo Jin, Zhaopeng Meng, Changming Sun, Hui Cui, and Ran Su. Ra-unet: A hybrid deep attention-aware network to extract liver and tumor in ct scans. Frontiers in Bioengineering and Biotechnology, 8:1471, 2020.

[32] Jo Schlemper, Ozan Oktay, Michiel Schaap, Mattias Heinrich, Bernhard Kainz, Ben Glocker, and Daniel Rueckert. Attention gated networks: Learning to leverage salient regions in medical images. Medical image analysis, 53:197–207, 2019.

[33] Dhiraj Maji, Prarthana Sigedar, and Munendra Singh. Attention res-unet with guided decoder for semantic segmentation of brain tumors. Biomedical Signal Processing and Control, 71:103077, 2022.

[34]Sharif, M.I.; Khan, M.A.; Alhussein, M.; Aurangzeb, K.; Raza, M. A decision support system for multimodal brain tumor classification using deep learning. Complex Intell. Syst. 2021, 8, 3007–3020.

[35] J. Stember and H. Shalu, "Deep reinforcement learning with automated label extraction from clinical reports accurately classifies 3D MRI brain volumes," arXiv (Cornell University), Jan. 2021, doi: https://doi.org/10.48550/arxiv.2106.09812.

[36] E. Jun, S. Jeong, D.-W. Heo, and H.-I. Suk, "Medical Transformer: Universal Brain Encoder for 3D MRI Analysis," arXiv (Cornell University), Jan. 2021, doi: https://doi.org/10.48550/arxiv.2104.13633.

[37] S. Tummala, S. Kadry, S. A. C. Bukhari, and H. T. Rauf, "Classification of Brain Tumor from Magnetic Resonance Imaging Using Vision Transformers Ensembling," Current Oncology, vol. 29, no. 10, pp. 7498–7511, Oct. 2022, doi: https://doi.org/10.3390/curroncol29100590.

[38] M. Aloraini, A. Khan, S. Aladhadh, S. Habib, M. F. Alsharekh, and M. Islam, "Combining the Transformer and Convolution for Effective Brain Tumor Classification Using MRI Images," Applied Sciences, vol. 13, no. 6, p. 3680, Jan. 2023, doi: https://doi.org/10.3390/app13063680

[39] W. Dai, H. Cui, X. Kan, Y. Guo, S. van Rooij, and C. Yang, "Transformer-Based Hierarchical Clustering for Brain Network Analysis," arXiv.org, May 06, 2023. https://arxiv.org/abs/2305.04142

Figure and tables

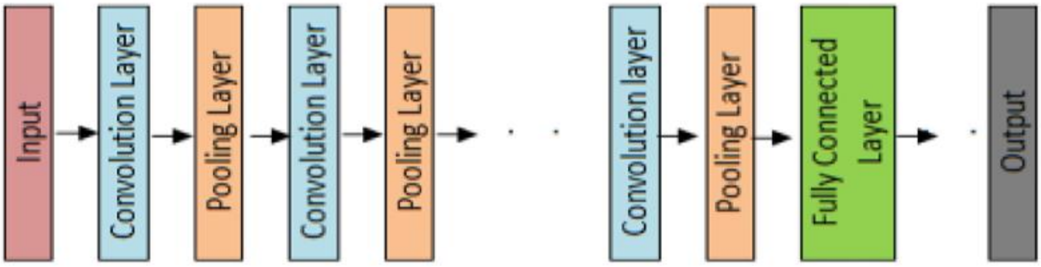

Fig. 1. Conceptual model of CNN

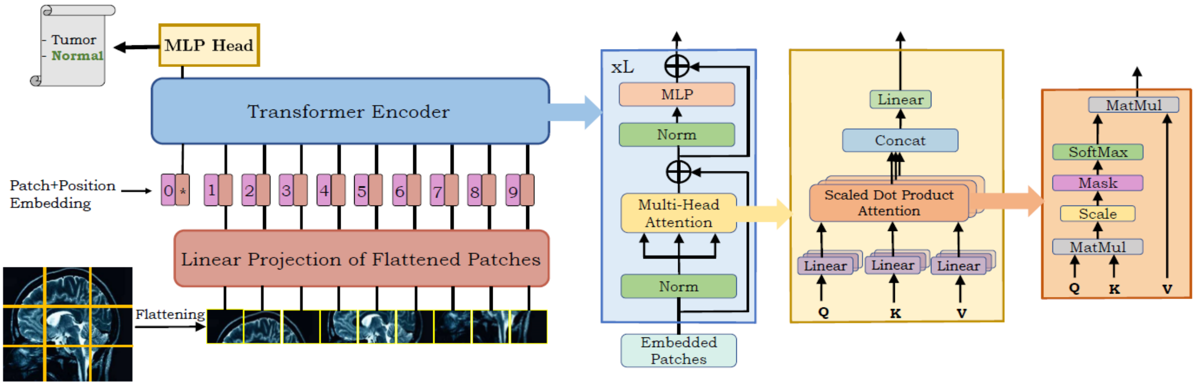

Fig. 2. Visual representation of the encoder block of the vision transducer (right) and its architecture (left). The transformer encoder processes the corrections that the vision transformer uses to generate the final classification output after displaying the corrections in the feature space after flattening.

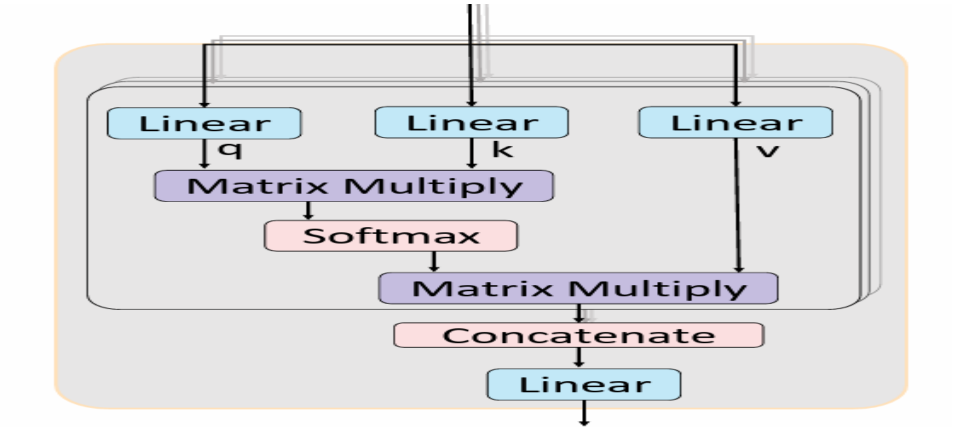

Fig.3. Illustration of self-attention in Transformer: Query, key, and value are denoted by q, k, and v, respectively; "Linear" refers to fully connected layers. Multi-vertex self-attention representations are pooled and passed through linear layers in order to design end-to-end sequences.

Table 1. Synopsis of the papers that there assessed classification of the brain MRI by deep learning.

Ref.	Classes	Features Extraction Methods and No. of Features	Classification methods	Metrics	Achievement	Limitation

HEBA M. AHMED et al.(2017)[7],	(a) Glioma (b) Meningioma (c)Pituitary tumor	Texture Feature Extraction with GLCM then selection features by GWO	Artificial Neural Network (ANN)	Accuracy: 0.9891 Sensitivity: 0.9970 Specificity: 0.9650 AUC: 0.9986	The computational complexity involved in optimization tasks such, as feature selection and fine tuning of DBN parameters can demand resources and time. By combining networks (ANN) with Grey Wolf Optimization (GWO) the integration leverages the distinct strengths of each method resulting in accurate and dependable identification of tumors, from MRI images.	The potential for overfitting in the classification model
A. Geetha, et al.(2020) [8]	Binary classification	GLCM, and the GLRM extracted fuzzy features then selection by GWO	Deep Belief Network	Accuracy: GW-DBN is 2.41%, 0.64%, 3.91%, and 5.48% better than FF-DBN, ABC-DBN, PS-DBN, and GA-DBN, respectively. Specificity: GW-DBN is 0.86% better than FF-DBN, 1.07% better than ABC-DBN, 35.63% better than PS-DBN, and 37.97% better than GA-DBN. False Positive Rate (FPR): GW-DBN is 97.76%, 96.24%, 97.75%, and 47.43% better than FF-DBN, ABC-DBN, PS-DBN, and GA-DBN, respectively.	The computational complexity involved in optimization procedures such, as feature selection and tuning DBN parameters can be quite demanding, necessitating processing capabilities and time.	The computational complexity involved in optimization procedures such, as feature selection and tuning DBN parameters can be quite demanding, necessitating processing capabilities and time.
Yakub Bhanothu et al.(2020) [9]	(a) Glioma (b) Meningioma (c)Pituitary tumor	The convolutional layers in VGG-16	Faster Region-based Convolutional Neural Network (Faster R-CNN)	precision Glioma: 75.18%. Meningioma: 89.45%. Pituitary Tumor: 68.18%. Overall precision Performance: 77.60%	VGG 16 has been effectively incorporated as the network, in the Faster R CNN framework, which has enhanced feature extraction leading to improved accuracy and reliability, in identifying and categorizing tumors.	The accuracy differs noticeably among types of tumors (, for instance accuracy for pituitary tumors) suggesting possible areas, for enhancement.
Srinath Kokkalla et al.(2021)[10]	(a) Glioma (b) Meningioma (c)Pituitary tumor	Inception ResNet v2 architecture	Deep Dense Network	Accuracy : 99.66 Precision: 99.60 Recall: 99.40 F1-score: 99.40	Enhancing the models capability to differentiate bettheen the three categories of brain is achieved through incorporating a dense network prior, to the softmax layer.	Concerns, about Overfitting; Even though dropout layers are employed to address overfitting issues the research recognizes that the abundance of features could still lead to overfitting concerns especially when dealing with datasets.
Aya M. Al-Zoghby et al.(2023)[11]	(a) Glioma (b) Meningioma (c)Pituitary	Integration of VGG-16, a pretrained convolutional	The Dual Convolution Tumor Network (DCTN)	Precision: 0.98 Recall : 0.99 F1-Score:0.98	The DCTN model shows performance, in categorizing brain tumors in MRI images suggesting its usefulness, in	Understanding learning models such, as DCTN can be tricky for healthcare professionals due, to their

	tumor	neural network (CNN),			different clinical settings and datasets.	complexity making it difficult to interpret the features or decision making processes involved thus limiting transparency and interpretability.
Prince Priya Malla et al.(2023)[12],	(a) Glioma (b) Meningioma (c)Pituitary tumor	Pre-trained DCNN architecture, VGG16	log-softmax layer	Accuracy: 98.93 Sensitivity: 98.68 Specificity: 99.13 Precision: 99.11	Detecting brain tumor types through MRI images with the suggested transfer learning technique.	Factors to take into account include how much pre trained models are used, assumptions regarding how features can be transferred the representativeness of the dataset and how effective data augmentation techniques are.
Elena Luisa et al.(2023)[13]	Binary Classification	ResNet-50	Convolutional Neural Network(CNN)	Accuracy: 94% Precision: 92% F1-Score: 90% Recall: 88%	The research project effectively created a network (CNN) system using a ResNet 50 design to identify brain tumors, in MRI scans.	The research made use of a dataset, from Kaggle, which might not completely represent the diversity and intricacy found in world clinical data.

Table 2. Synopsis of the papers that there assessed classification of the brain MRI by transformer.

Ref.	Classes	Features Extraction Methods and No. of Features	Classification Methods	Metrics	Achievement	Limitations
Joseph, et al.(2021) [35]	Normal , metastasis containing, tumor	- SBERT extracts features from the text data in the clinical reports to predict class labels for MRI brain scans. - Spatial features from 3D MRI brain image volumes are extracted using 3D convolutions.	-SBERT (sentence bidirectional encoder representations from transformers). -Deep-Q Network (DQN) combined with reinforcement learning (RL)	-Automated Label Extraction: 100% accuracy -3D Image Classification with RL and DQN: 92% accuracy -Supervised Deep Learning Classification: 66% accuracy	Successfully extracted class labels from reports, through Automated Label Extraction using SBERT. This study builds on work in 2D image categorization by incorporating reinforcement learning (RL) with a Deep Q Network (DQN) for analyzing 3D image volumes. The effectiveness of RL in training with data sets showcases its capacity for learning applications particularly in scenarios, with small training datasets.	- Overfitting in Supervised Approach - Limited Testing Data - Dependency on Clinical Reports
Eunji Jun, et al.(2021) [36]	Meningiomas Gliomas Pituitary	The convolutional encoder serves as the backbone for extracting features from 2D image slices of 3D MRI scans	Feed-forward Neural Networks	AUC 0.8347 ± 00072	A strong framework has been created to efficiently transfer expertise from pre training, to medical imaging assignments enhancing effectiveness and flexibility in the analysis of 3D MRI scans.	The effectiveness of the model could vary depending on the characteristics of the data or the different populations it is applied to.
Sudhakar, et al.(2022)	Meningiomas Gliomas	ViT architecture.	ViT model's(B/16, B/32, L/16, and L/32)	accuracy of L/32 is 98.2% and accuracy of(B/16, B/32, L/16,) is 98.7%	Showcasing the effectiveness of combining multiple ViT architectures for brain tumor classification	The accuracy of the Vision Transformer (ViT) models was measured at a 384x384 resolution showing a difference, in

Ref.	Classes	Features Extraction Methods and No. of Features	Classification Methods	Metrics		Achievement	Limitations
[37]	Pituitary					from T1 MRI images.	performance compared to the 224x224 resolution. This variation could be limiting in scenarios where using high resolution imagess difficult or not feasible.
Mohammed ,et al.(2023) [38].	Meningiomas Gliomas Pituitary	The CNN is utilized to extract features to a location while the transformer utilizes an attention mechanism to capture features on a scale.	intelligent merge module (IMM).	Data1 Accuracy 96.75 Precision 0.967 Recall 0.970 F1-Score 0.968	Data 2 Accuracy 99.10 Precision 0.987 Recall 0.99 F1-Score 0.987	A new model has been created that merges local and global data extraction techniques to classify brain tumors. Through the use of both the Feature Fusion Module (FFM) and Information Fusion Module (IFM) this model shows results, than approaches highlighting the effectiveness of extracting local and global features simultaneously.	The challenges and expenses involved in setting up and training FFM and IFM modules are significant. Moreover the models adaptability, to datasets or imaging methods may not have been fully assessed, potentially impacting its effectiveness in a range of situations.
Thei Dai, et al.(2023) [39].	Binary classification.	Understanding how to utilize node embeddings and attention matrices, with transformer encoders is crucial for extracting features from adjacency matrices. Additionally the application of clustering methods plays a role in pinpointing functional modules within the brain network, which in turn facilitates feature extraction, for classifying brain networks.	Transformer-based Hierarchical Clustering model	Dataset: ABIDE AUROC: 79.76 Accuracy: 70.6 Dataset: ABCD AUROC: 96.2 Accuracy: 89.4		Increased Accuracy and Reduced Runtime Complexity of MRI classification	Choosing the hyperparameters, such, as cluster thresholds, attention mechanisms and the number of layers can greatly influence how the model performs. To tune these hyperparameters effectively it might require an amount of computational resources and experimentation.