RESEARCH ON NEURAL NETWORK ANALYSIS OF MRI IMAGES OF THE BRAIN FOR THE DIAGNOSIS OF BRAIN DISEASES

Tekere Richard [1], Spitsyn V.G [2]
[1] TPU, SCSR, Gr. 8VM02, e-mail: tekere@tpu.ru
[2] TPU, SCSR, Doctor of Technical Sciences
Professor., e-mail: spvg@tpu.ru

Abstract

This study presents a deep learning-based approach for MRI brain disease diagnosis using TransUNet, integrating CNNs, Vision Transformers, and Explainable AI techniques. The proposed system enhances segmentation accuracy and interpretability, bridging the gap between AI research and clinical application.

Keywords: Brain Tumor Segmentation, MRI Analysis, Deep Learning, Vision Transformers, Explainable AI, TransUNet.

Introduction

Brain diseases, including tumors and neurodegenerative disorders, remain a major challenge in modern medicine, requiring early and precise diagnosis for effective treatment. Traditional manual analysis of MRI scans by radiologists is time-consuming, prone to human error, and lacks consistency. The application of neural networks in medical imaging has shown promising results in improving diagnostic accuracy and efficiency [1]. However, the lack of transparency in deep learning models often limits their adoption in clinical practice. Therefore, developing an interpretable neural network-based diagnostic tool is crucial for integrating AI into healthcare.

Recent research efforts have demonstrated the effectiveness of Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) in medical imaging [4]. CNN-based models excel at capturing local spatial features, while ViTs excel in processing global dependencies across an image. The combination of these architectures in TransUNet has shown promising results in brain tumor segmentation tasks, outperforming traditional CNN-based models [1]. Despite these advancements, the black-box nature of deep learning models presents a significant barrier to clinical adoption.

To address this issue, Explainable AI (XAI) techniques, such as Grad-CAM and SHAP, have been introduced to improve the transparency of neural network decisions [2]. By integrating XAI methods, we can generate visual explanations for segmentation results, providing clinicians with greater trust in AI-assisted diagnosis. This paper explores the integration of TransUNet and XAI techniques into an interactive tool for MRI-based brain tumor segmentation [3].

This study aims to develop a neural network-based diagnostic system that integrates CNNs, Vision Transformers, and Explainable AI for the analysis of MRI images of the brain for disease diagnosis. The system will provide automated disease detection and interpretability features, enabling radiologists to make informed clinical decisions. We focus on the BraTS 2021 Task 1 dataset, a benchmark dataset for brain tumor segmentation, to evaluate the proposed methodology. The system is deployed as a web-based application, allowing radiologists to upload MRI scans, perform automated segmentation, visualize explainability heatmaps, and export results. The goal is to provide a clinically useful tool that combines high accuracy with interpretability, thereby bridging the gap between AI research and practical medical applications.

Proposed Methodology

The proposed methodology for MRI-based brain tumor segmentation using TransUNet is illustrated in Fig.1.The proposed transUnet will be built using Pytorch. It consists of multiple stages, including preprocessing, feature extraction, segmentation, and explainability.

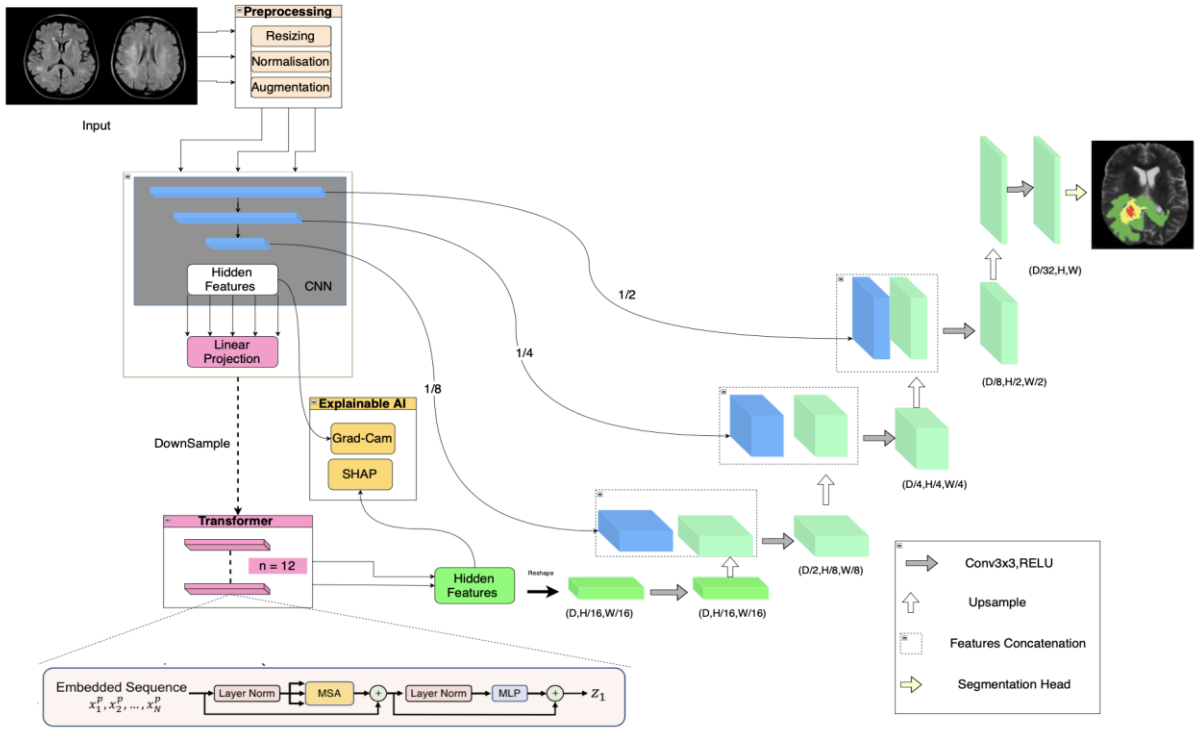

Fig. 1. Schematic representation of the proposed TransUNet architecture for diagnosis MRI brain diseases

Preprocessing

The input MRI scans undergo preprocessing steps, including:

1. Resizing: Standardizing all images to a fixed dimension to match the model's input requirements.

2. Normalization: Adjusting pixel intensity values to a uniform range to improve model performance.

3. Augmentation: Applying transformations such as rotation and contrast adjustments to enhance generalization.

Feature Extraction (CNN Encoder & Transformer Encoder)

1. CNN Encoder: Extracts local spatial features from the MRI scan. Feature maps are extracted using convolutional layers, which apply filters to input images. The transformation at each layer is given by:

$$F_{l+1} = f(W_l \star F_l + b_l), \quad (1)$$

where F_l is the feature map, W_l and b_l are weights and biases, f is the activation function (ReLU) and $\star$ represents the convolution operation, which slides the filter over the input to extract spatial patterns.

2. Transformer Encoder: Captures global dependencies in the image, processing features as embedded sequences. The Transformer encoder captures global dependencies by computing multi-head self-attention (MSA) across image patches The self-attention mechanism is formulated as:

$$MSA(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \quad (2)$$

where Q, K, V are query, key, and value matrices derived from input features, and d_k is the dimension of the key vectors. This mechanism enables the model to focus on important image regions by assigning higher attention weights.

After applying Multi-Head Self-Attention (MSA) using eq. (2), the Transformer encoder processes features using a Multilayer Perceptron (MLP) block, which consists of two fully connected layers with a nonlinear activation function:

$$MLP(X) = \sigma\big(W_2\big(\mathrm{ReLU}(W_1X + b_1)\big) + b_2\big) \quad (3)$$

where X is the input feature representation, W_1,W_2 are weight matrices, b_1,b_2 are biases, and σ is an activation function such as Softmax or RELU. Each Transformer layer updates the input features using the following sequence:

$$z_l' = MSA\big(LN(z_{l-1})\big) + z_{l-1} \quad (4)$$

$$z_l = MLP\big(LN(z_l')\big) + z_l' \quad (5)$$

where LN (·) represents Layer Normalization, MSA (·) applies self-attention across image patches, and MLP (·) is a two-layer feedforward network that refines feature representations. The MSA block learns contextual relationships between different image regions, while the MLP block enhances learned embeddings before passing them to the next Transformer layer.

3. Feature Concatenation: The extracted CNN and Transformer features are merged before being passed to the decoder.

Segmentation (CNN Decoder & Segmentation Head)

1. CNN Decoder: Reconstructs the segmentation mask by progressively upsampling features.

2. Segmentation Head: The final layer applies a 1×1 convolution to generate a pixel-wise classification map.

Explainability Module (Grad-CAM & SHAP)

1. Grad-CAM (Gradient-weighted Class Activation Mapping): Applied to the last layer of the CNN encoder to generate heatmaps showing important regions for segmentation. To interpret the CNN encoder's decisions, Gradient-weighted Class Activation Mapping (Grad-CAM) generates heatmaps highlighting important image regions. The activation map is computed as:

$$L^c_{Grad-CAM} = ReLU(\textstyle\sum_k \alpha_k^c A^k), \quad (6)$$

where α_k^c represents the importance weights, computed as:

$$\alpha_k^c = \frac{1}{Z}\textstyle\sum_i \sum_j \frac{\partial y_c}{\partial A_{ij}^k}, \quad (7)$$

Where, A^k represents feature maps, and $\frac{\partial y_c}{\partial A_{ij}^k}$ denotes the gradient of the class score with respect to feature activations.

2. SHAP (Shapley Additive Explanations): Applied to the Transformer Encoder output, identifying which feature patches contributed most to the segmentation decision. To further interpret the Transformer encoder's contribution to segmentation, we apply Shapley Additive Explanations (SHAP) to quantify feature importance. The Shapley value for each feature is calculated as:

$$\phi_i = \textstyle\sum_{s \subseteq F/\{i\}} \frac{|s|!(|F|-|s|-1)!}{|F|!}[f(S \cup \{i\}) - f(S)], \quad (8)$$

where ϕ_i represents the contribution of feature i to the prediction, and f(S) is the model output for a given subset of features S. SHAP helps identify the most influential image regions for segmentation.

Web-Based Interactive Tool

A Django-based web application will be developed to provide:

1. MRI Upload Interface – Users can upload MRI scans in DICOM/NIfTI format.
2. Automated Analysis – The uploaded image is processed using the deep learning model.
3. Explainability Module – AI-generated explanations help interpret results.
4. User Visualization – Disease markers and heatmaps are overlaid on MRI scans.
5. Export Functionality – Results can be saved in PNG/PDF format for further review.

This methodology ensures accurate segmentation while enhancing interpretability, allowing radiologists to visualize tumor locations and understand the model's decision-making process.

Results

Since the model is under development, the model is expected to achieve a Dice Similarity Coefficient (DSC) above 0.85, ensuring precise tumor segmentation. Performance will be evaluated on the BraTS 2021 Task 1 dataset using standard metrics. Regarding explainability, Grad-CAM heatmaps will highlight crucial regions in MRI scans that influence segmentation outcomes, providing intuitive visual feedback for radiologists. SHAP values will quantify the importance of different patches within the Transformer encoder, offering additional interpretability regarding how various MRI features contribute to tumor identification. These insights will help build trust in AI-assisted medical diagnosis and facilitate the clinical adoption of deep learning-based segmentation models.

If preliminary experiments are conducted before submission, initial classification and segmentation results will be presented.

Conclusion

This study proposes a deep learning-based MRI analysis system that integrates CNNs, Transformers, and Explainable AI techniques for brain disease diagnosis. The system is designed to improve the accuracy and efficiency of MRI-based disease detection while enhancing model interpretability through the use of SHAP and Grad-CAM. Furthermore, it provides a web-based interactive tool that enables radiologists to perform automated segmentation, analyze explainability visualizations, and export results for further assessment.

Future work will focus on extending the dataset to include more diverse brain disease cases to improve generalization. Additionally, optimizing the deep learning model for faster inference will be prioritized to ensure real-time usability in clinical settings. Finally, conducting clinical trials will be essential to evaluate the system’s effectiveness in real-world medical environments.

This study bridges the gap between cutting-edge AI research and clinical applications, ensuring that deep learning-driven medical diagnostics remain both accurate and explainable.

References

1. Chen J., Lu Y., Yu Q., Luo X., Adeli E., Wang Y., Chang E.I.C., Xu Y. TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation. In: Medical Image Analysis. – 2021.

2. Selvaraju R. R., Cogswell M., Das A., Vedantam R., Parikh D., Batra D. Grad-CAM: Visual Explanations from Deep Networks. In: IEEE International Conference on Computer Vision. – 2017.

3. Bakas S., Reyes M., Jakab A., Bauer S., Rempfler M., Crimi A., Shinohara R. T., Berger C., Ha S. M., Rozycki M., Kirschke J. Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge. In: IEEE Transactions on Medical Imaging. – 2018.

4. Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: arXiv preprint arXiv:2010.11929. – 2020.

5. Rafii M.S., Walsh S.J., Little B.C., Fogel A.L., Thompson P.M., Schwartz J.B. Artificial intelligence in brain MRI analysis of Alzheimer's disease over the past 12 years: A systematic review. Ageing Research Reviews. – 2022. DOI: 10.1016/j.arr.2022.101614.

6. Moradi E., Pepe A., Gaser C., Huttunen H., Tohka J. A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to Alzheimer's disease. NeuroImage. – 2019. DOI: 10.1016/j.neuroimage.2019.01.031.