Mutltimodal MRI Brain Tumor Segmentation using 3D Attention U-Net with Dense Encoder Blocks and Residual Decoder Blocks

Tewodros Megabiaw Tassew (ted.meg1234@mail.nwpu.edu.cn)  
Northwestern Polytechnical University

Betelihem Asfaw Ashamo (asfawbetelihem@gmail.com)  
Northwestern Polytechnical University

Xuan Nie  
Northwestern Polytechnical University

Method Article

Keywords: Brain tumor segmentation, MRI images, Attention U-Net, Dense block, Residual block

Posted Date: March 27th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2717573/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

Medical image segmentation is essential for disease diagnosis and for supporting medical decision systems. Automatic segmentation of brain tumors from Magnetic Resonance Imaging (MRI) is crucial for treatment planning and timely diagnosis. Due to the enormous amount of data that MRI provides as well as the variability in the location and size of the tumor, automatic segmentation is a difficult process. Consequently, a current outstanding problem in the field of deep learning-based medical image analysis is the development of an accurate and trustworthy way to separate the tumorous region from healthy tissues. In this paper, we propose a novel 3D Attention U-Net with dense encoder blocks and residual decoder blocks, which combines the benefits of both DenseNet and ResNet. Dense blocks with transition layers help to strengthen feature propagation, reduce vanishing gradient, and increase the receptive field. Because each layer receives feature maps from all previous layers, the network can be made thinner and more compact. To make predictions, it considers both low-level and high-level features at the same time. In addition, shortcut connections between the residual network are used to preserve low-level features at each level. As part of the proposed architecture, skip connections between dense and residual blocks are utilized along with an attention layer to speed up the training process. The proposed architecture was trained and validated using BraTS 2020 dataset, it showed promising results with dice scores of 0.866, 0.889, and 0.828 for the tumor core (TC), whole tumor (WT), and enhancing tumor (ET), respectively. In comparison to the original 3D U-Net, our approach performs better. According to the findings of our experiment, our approach is a competitive automatic brain tumor segmentation method when compared to some state-of-the-art techniques.

1. Introduction

Brain tumors are a serious medical condition that can have devastating effects on the nervous system and surrounding healthy brain tissue. They are divided into two main categories: primary and metastatic. Primary brain tumors originate from the cells of the brain or its immediate vicinity, while metastatic ones form elsewhere in the body then spread to the brain through circulation. Primary tumors can be classified further as glial, composed of glial cells, or non-glial, formed on or in other structures such as nerves, blood vessels, and glands; these may be benign (non-cancerous) or malignant (cancerous). Treatment for both types varies depending on how advanced they are but typically involves surgery followed by radiation therapy or chemotherapy depending upon individual circumstances. In the United States alone, 700,000 people had primary brain tumors, with over 88,000 diagnosed by 2022. According to the World Health Organization, cancer causes 9 million deaths globally each year, accounting for 22 percent of all chronic diseases and ranking second only to cardiovascular disease, which causes 17.9 million deaths each year. Primary malignant brain tumors, for example, account for 1.4 percent of all new cancer diagnoses in the United States each year. Primary malignant brain tumors have a global mortality rate of 2.0/100,000 in women and 2.8/100,000 in men (Ellison, 2020).

Medical imaging is an essential tool in the diagnosis and treatment of tumors. It allows for accurate tumor segmentation, assessment of tumor size and shape, and evaluation of response to treatments such
as radiation or chemotherapy. The most commonly used medical imaging techniques are computed tomography (CT) scans, ultrasound examinations, X-ray examinations, and magnetic resonance imaging (MRI). These technologies allow healthcare professionals to observe the structure of tissues within a patient's body in order to accurately diagnose tumors or assess how well a treatment is working. Medical imaging plays an important role in providing early detection of cancerous lesions which can lead to better prognosis outcomes for patients with tumors (Raza et al., 2023). In clinical medicine, MRI has become the most widely used imaging examination for brain tumor diagnosis, because of its sharpness and high tissue resolution, ability to produce images of the brain's tissues in clear contrast, and versatility in allowing different parameters to be set in order to obtain specific anatomical information (Raza et al., 2023; Alagarsamy et al., 2020). It is a safe and advanced imaging procedure that makes use of nuclear magnetic resonance to produce a reliable depiction of internal body structure and tissues without causing harm to the organ being imaged or causing high levels of ionization or radiation effect. It creates an accurate portrayal of internal body structure and tissues by combining strong magnetic waves ranging from 1.5 T to 3 T with radiofrequency waves (Azhari et al., 2014; Tanneedi et al., 2017). The use of MRI for glioma diagnoses allows physicians access to detailed information about the cells within a patient's brain by offering up sub-regions such as peritumoral edema, necrotic core, enhancing tumor core, and non-enhancing tumor core depending on factors such as degree of invasion or prognosis. This technology also helps healthcare professionals determine if any additional treatments are necessary after the initial diagnosis has been made since it can provide further insight into any changes that may have occurred over time regarding the tumor's size or shape prior to beginning treatment (Alagarsamy et al., 2020; Crimi et al., 2019; Ghaffari et al., 2020).

Multimodal MRI scans are a powerful tool for diagnosing and monitoring diseases of the brain. By combining four different modalities, T1 weighted (T1), T1 weighted with post-contrast (T1Gd), T2 weighted (T2), and Fluid Attenuated Inversion Recovery (FLAIR) scans, it is possible to gain an accurate representation of the various histological sub-regions within the brain (Ghaffari et al., 2020; Zhang et al., 2011). The purpose of each modality can be broadly understood based on their distinct intensity distributions (Zhang et al., 2020). For example, T1w is mainly used to measure healthy tissues. While T2w has a bright tumor region visible on scan results, the FLAIR scan helps distinguish edema from cerebrospinal fluid (CSF) (Bauer et al., 2013). Meanwhile, both the whole tumor area as well as its core can be seen clearly when viewing axial slices through both a FLAIR and T2 scan, respectively. Additionally, a post-contrast enhanced view via a T1Gd allows for the visualization of enhancing tumors. This combination provides clinicians with valuable insight into disease states that would otherwise remain undetected or overlooked.

An example demonstrating these four MRI modalities is shown in Figure 1, where ground truth segmentation was applied to patient data from the BraTS2020 training dataset. This demonstrates how multimodal MRIs provide invaluable information regarding tissue structure, which helps build an
understanding of normal physiology versus pathological conditions. In conclusion, multimodal MRIs are essential tools used by healthcare professionals across many disciplines due to their ability accurately reflect differences between healthy tissues and diseased states without resorting to invasive procedures such as biopsy or surgery.

Automation plays an important role when it comes to accurately identifying areas affected by cancerous cells without damaging healthy parts of the body during surgery. Manual segmentation of brain tumors from MRI images is a time-consuming process that can be easily influenced by the subjective judgement of experts. Therefore, automated and semi-automated approaches have been developed over the last two decades to improve accuracy and reduce the time needed for this task. The challenge with automated segmentation lies in its complexity as tumorous cells may appear anywhere inside the brain tissues with varying size, shape or appearance which makes it difficult to distinguish them from normal tissue boundaries due to poor contrast (Al-qazzaz, 2020; Baid et al., 2020). Moreover, different imaging modalities such as CT scans or Magnetic Resonance Imaging (MRI) provide different types of information about a tumor’s location within the brain which requires specialized algorithms tailored specifically for each type. This has led researchers towards developing more sophisticated techniques such as deep learning models that incorporate prior knowledge into their predictions, allowing radiologists to speed up their workflow while achieving consistent results across multiple cases (Liu et al., 2022).

2. Related Works

The early detection and treatment of a brain tumor are essential because it is a potentially fatal condition. In the past, traditional machine learning methods that relied on hand-crafted feature engineering were used to acquire representations from brain images. However, these methods are not sufficient for accurate brain tumor segmentation (BTS) due to their limited ability in capturing local information of the tumor regions (Raza et al., 2023). Recent studies show that there are currently three different types of brain tumor segmentation algorithms: classical approaches, machine learning approaches (excluding deep learning approaches), and deep learning approaches (Tie et al., 2021).

The most popular traditional approaches are the edge detection method (Caselles et al., 1993; Muthukrishnan and Radha, 2012), threshold method (Prastawa et al., 2004; Stadlbauer et al., 2004), region growth method (Gibbs et al., 1996; Weglinski and Fabijanska, 2011), watershed approach (Lin et al., 2006; Maiti and Chakraborty, 2013), and level set approach (Menze et al., 2010; Prastawa et al., 2003). These approaches are rapid and inexpensive to deploy, but they have limits due to their inability to generate good results in difficult cases, such as large or irregularly shaped tumors. Edge detection works by extracting boundary information from images while threshold involves setting a certain intensity value that divides objects from background pixels; both have limited accuracy when dealing with noisy or blurry images. Region growing begins at a point within the object and expands outward until it reaches the boundaries defined by user input parameters; however, this technique does not always guarantee accurate results because it may miss small regions within an image or include parts outside its borders due to noise levels in the image data. Watersheds use gradient magnitude maps in combination with markers
manually placed at the center of each object. Level sets employ partial differential equations that evolve contours over time according to how they interact with edges detected using gradient operators. Both require more computational power than simpler techniques like edge/thresholding yet still lack robustness when faced against challenging scenarios where multiple objects exist close together sharing similar intensities across their boundaries making them difficult distinguishable from each other.

Machine learning methods are increasingly being utilized for the purpose of object classification and detection. These machine learning algorithms use training data to autonomously learn the features of a given dataset, allowing them to accurately classify objects within it. This has been especially useful in medical imaging applications such as brain tumor segmentation, where researchers have used machine learning methods like support vector machines (SVM) (Bauer et al., 2011; Ruan et al., 2007), clustering-based approaches (Stupp et al., 2017; Narayanan et al., 2019), conditional random fields (CRF) (Meier et al., 2014, 2015) and random forest techniques (RF) (Tustison et al., 2015; Breiman, 2001). Bauer et al (Bauer et al., 2011), for example, developed a classification method based on hierarchical SVM in conjunction with a CRF-based technique to segment a brain tumor from 3D MRI images. And Ruan et al. (Ruan et al., 2007) employed support vector machines (SVMs) to detect the abnormal region and then used multi-scales to extract the brain tumor field. Spatially Constrained Fish School Optimization (SCFSO) and Interval type-II Fuzzy Logic System (IT2FLS) algorithms were proposed by Stupp et al. (Stupp et al., 2017) for segmenting tumor and non-tumor regions in MR brain images. The segmentation results obtained by applying IT2FLS to the subsequent clustering assignment following the use of SCFOS were impressive, taking less processing time while delivering good results.

For segmenting brain tumors in multi-modal MR images, Narayanan et al. (Narayanan et al., 2019) combined particle swarm optimization (PSO) and bacterial foraging optimization (BFO) with an upgraded clustering method based on fuzzy c means. Whereas Tustison et al. (Tustison et al., 2015) suggested a supervised brain tumor segmentation approach based on random forest (RF) and Markov random field that produced good performance when applied to multimodal MRI scans, Meier et al. (Meier et al., 2015) employed decision forest for classification and conditional random field (CRF) for spatial regularization in brain tumor segmentation. The main benefit of using these types of machine learning algorithms is that they require minimal human intervention; once trained on a given dataset they can be applied quickly and efficiently without requiring additional input from operators or clinicians. Furthermore, their accuracy tends to improve over time as more training data becomes an important factor when dealing with rare diseases or conditions which may not have sufficient datasets readily available for analysis at first instance.

In recent years, deep learning which is a particular kind of machine learning has made remarkable achievements in many fields, such as pattern recognition, image segmentation, image detection and image classification (Zhang et al., 2019, 2021; Xiang et al., 2021). It has reentered the spotlight ever since AlexNet won the ImageNet competition in 2012 (Krizhevsky et al., 2012; Deng et al., 2009). Many
Researchers have applied deep learning methods for medical imaging applications such as the automatic segmentation of brain tumors from MRI scans. Deep Learning models have been shown to outperform traditional machine learning algorithms when it comes to accuracy rate for automated BTS segmentation tasks with minimal human intervention required during the training phase itself.

The early application of deep learning in brain tumor segmentation is mainly to modify the convolution network. It is able to extract features from volumetric scans using convolutional neural networks (CNNs). CNNs can capture both global and local information about the scanned image, allowing them to accurately detect subtle changes in tissue structure. Convolutional neural networks (CNNs) have revolutionized the field of deep learning and made tremendous progress in areas such as computer vision, natural language processing, and robotics. CNNs are a type of artificial neural network for automatically learning high-dimensional hierarchical features that use convolution operations to extract features from an input image or video. This has enabled researchers to create several specialized deep learning models such as ResNet (He et al., 2016), VGG-Net (Szegedy et al., 2015), and Inception networks (Szegedy et al., 2017) for various tasks. These architectures can be used together with open-source frameworks like TensorFlow and PyTorch which makes development easier in academic fields like computer vision where semantic segmentation is used to classify objects within images or videos accurately. Fully convolutional networks (FCN) enable CNNs to label each pixel using a straightforward upsampling technique which has proven invaluable when it comes to recognizing patterns at multiple resolutions without needing additional training data for different scales or sizes of inputs (Long et al., 2015).

Because of the rapid progress of machine and deep learning techniques, we are now able to generate 3D volumetric data from 2D brain imaging slices in a timely, accurate, and cost-effective manner. This enables the adoption of more complicated models like U-Net, VNet, and other 3D models that can extract a greater number of features than typical 2D models. U-Net is one of the most widely used deep learning architectures for biomedical image segmentation. It was first introduced by Ronneberger et al. (Ronneberger et al., 2015) and has since gained popularity because of its ability to deliver good segmentation results despite limited training data. The U-Net architecture is a powerful tool for segmentation tasks, particularly in the medical field. Its left contracting (encoder) part is responsible for feature extraction while its right expanding (decoder) part produces an output and maps it back into raw image pixels. Skip connections are essential to this model, as they allow the extracted features from the encoder block to be passed on to the decoder block so that tumors can be detected accurately. In recent years, different studies have used combinations of skip connections with models such as U-Net++ (Zongwei et al., 2018) and Net3+ (Huimin et al., 2020).

Furthermore, attention models (Prajit et al., 2019) have also been actively used in vision tasks due to their potential accuracy increase when compared with traditional methods. Attention models filter out irrelevant features from input images which save computational power and improve performance; thus making them a popular choice among researchers using U-Nets in medical applications. Several studies have been undertaken in the medical industry, particularly with U-Net, to boost the accuracy of the model with attention models filtering away unimportant elements from the input images to conserve
computational power and accelerate the model. For instance, the authors of (Oktay et al., 2018) developed a unique attention gating approach based on the U-Net (Olaf et al., 2015) architecture. In their research effort, the experts found that the attention module suppresses the unnecessary areas of the input image, while accentuating the salient features. They added that an attention module-enhanced model can operate with little additional computational burden while improving the model’s sensitivity and predictive power.

The use of 2D or 3D convolutions during the training of a deep CNN model is essential for accurate and effective segmentation. U-Net, Res-Net, and DenseNets (Simon et al., 2017) were previously proposed, but they only utilized two-dimensional input images thus missing out on additional contextual information between slices. 2D convolutional networks are limited in their ability to capture spatial information from medical images, while 3D convolutions require more memory and computational resources. To address this issue, W. Chen et al (Chen et al., 2019) proposed a separable 3D U-Net model that uses separate 2D and 1D layers to learn both spatial features (2D) as well as temporal features (1D). This approach allows for full utilization of all three dimensions within the brain volume without overburdening memory requirements or taking up too much computing power. In order to train their proposed approach for each orthogonal view (axial, sagittal, and coronal) individually, the authors added separable temporal convolution to the residual inception model. For improved performance, they also used a multi-view fusion approach to merge all of the convolutional outputs. In terms of efficiency, the suggested model performed well on the BraTS 2018 test dataset. For patients with brain cancer or other illnesses that generate tumor growths inside the skull cavity, accurate brain image segmentation models are crucial for early tumor detection and enhancing survival rates. Local and global characteristics are vital for making judgments regarding tumor identification; however, gradients of low-level features approach zero as we move deeper into this process, making it impossible to identify borders or edges effectively without utilizing more sophisticated methods such as deep learning techniques like those utilized by these researchers in their study.

The residual ResNet with a residual block was proposed by He et al. (He et al., 2016) to simplify neural network optimization and address the degradation problem in deep learning architectures. This type of network has been widely applied for image classification, recognition and segmentation tasks. Myronenko et al. (Myronenko, 2019) further improved this architecture by adding a variational auto-encoder (VAE) branch based on 3D U-Net while replacing the original U-Net convolutional blocks with residual blocks. Huang et al. (Huang et al., 2017) proposed Dense Convolutional Networks (DenseNets), which are designed such that all layers are directly connected while ensuring maximum information transmission between them; thereby allowing more effective feature propagation, alleviating vanishing gradients issues, and reducing parameter numbers at the same time. Subsequently, many researchers have combined DenseNets with U-nets for medical image segmentation tasks as well due to their effectiveness in these areas of application.

MDU-Net (Zhang et al., 2019) is an improved version of U-net that uses three different multi-scale dense connections in the encoder, decoder and connection between them. This type of network was first
proposed by Wang et al. (Wang et al., 2019) to segment retinal vessels and achieved good results. Building on this success, Ziang et al. (Ziang et al., 2020) combined Inception Residuals with dense connections to further improve the performance of UNet for medical image segmentation tasks such as tumor detection or brain tissue classification.

By combining the advantages of DenseNet and ResNet, Tie et al. (Tie et al., 2021) proposed a new 3D U-Net with dense encoder blocks and residual decoder blocks for segmenting the brain tumor multi-class 3D MR images and achieved better results. Raza et al. (Raza et al., 2023) proposed a combined model of deep residual network and U-Net model (dResU-Net) to make predictions using both low-level and high-level information at the same time. Moreover, shortcut connections between residual networks are used to maintain low-level characteristics at each level and to enhance the training in the brain tumor segmentation process. According to results that were obtained by comparing the works, dResU-Net has greatly enhanced the segmentation performance of brain tumor sub-regions.

Due to the enormous variability of MRI brain tumor images and the small proportion of brain tumor, segmentation is an extremely tough task. Although deep learning has produced significant improvement and impressive results in segmentation with top results of dice scores 0.8660, 0.8357, and 0.8004 for whole tumor, tumor core, and enhancing tumor core respectively on the BraTS2020 validation data set, there is no doubt that this segmentation accuracy could not be used for clinical diagnosis. MRI brain tumor segmentation still has to improve a lot until completely addressed. This paper proposes a new network architecture based on 3D Attention U-Net to solve the accuracy problem of MRI brain tumor segmentation.

3. Methodology

This section discusses the dataset's details, preprocessing procedures, and the proposed method's implementation in detail. The proposed 3D Attention U-Net architecture is also covered in this part, along with detail on the loss functions that were used during the training stage.

3.1. Dataset

The dataset that we used to train our model is taken from "Medical Image Computing and Computer-Assisted Intervention (MICCAI) Multimodal Brain Tumor Segmentation Challenge (BraTS) 2020" which is publicly available in BraTS 2020 challenge dataset. It was gathered by medical experts from the “University of Pennsylvania and UPenn's Center for Biomedical Image Computing and Analysis (CBICA)” (Menze, 2015).

It is divided into training and validation datasets which are manually annotated by neuroradiologists. It includes 3D MRI brain scans from 369 patients with gliomas, 76 of them had LGG, while the rest images are from patients with HGG. Each 3D scan consists of 155 slices, with each image measuring 240 x 240 dimensions. There are four main types of labels used to describe the different areas of a brain tumor including Label 0, Label 1, Label 2, and Label 4. Label 0 is reserved for non-tumor areas such as
background or normal tissue area while Labels 1 and 4 refer to neurotic (NCR) or non-enhancing core (NET), edema (ED), respectively as well as enhancing tumor cores (ET). The regions labeled greater than 0 are referred to collectively as “whole tumor” regions with those labeled specifically 1 and 4 is called “tumor core” regions containing the ET region.

3.2. Pre-Processing

The BraTS 2020 dataset was preprocessed by the competition's organizers before it was made publicly available. The images were all co-registered to the same anatomical template, interpolated to the same resolution (1 mm$^3$), and have isotropic resolution. Despite the fact that all MRI images have been preprocessed, the brightness of the images vary substantially. As a result, we normalized each 3D MR modality image using a Min-Max normalizer to remove the impact of image intensity change on prediction accuracy. By doing so, all features will be converted into a range [0,1], with 0 and 1 serving as the minimum and maximum values for each feature/variable. It is calculated by the following formula:

$$
x_{\text{scaled}} = \frac{x - \min(x)}{\max(x) - \min(x)}
$$

where $x_{\text{scaled}}$ is the new value of each entry in the data and $x$ is the old value of each entry in data. $\max(x)$ and $\min(x)$ are the maximum and minimum values of the features respectively. Due to memory constraints, all MR images and segmentation masks were reduced to 120 x 120 x 96 dimensions from 240 x 240 x 155.

Nearest-neighbor interpolation was used to resize both the images and the masks. When resizing segmentation masks, it is necessary to apply ‘nearest neighbor’ interpolation. This is because other techniques, such as Bilinear interpolation, consider the closest 2x2 neighborhood of known pixel values surrounding a pixel. It then takes a weighted average of those pixels to get the final interpolated value, which can be a float value. This creates a smoother image than nearest-neighbor interpolation, but it results in full black images for some slices having segmentation masks.

Nearest-neighbor does not really interpolate the pixels. Instead, it looks for the closest one in the source image that matches the location of the target image. It’s the fastest interpolation method, and requires the least amount of processing time, and allows for sharper detail. The biggest drawback, however, is that the resulting image may contain jagged edges. But it is suitable for this task as it wont affect the segmentation masks. All four modalities were stacked together to take advantage of the entire information given in four sequences. At the training time, training examples with dimensions of 120x120x96x4 (where 4 denotes the four modalities, i.e., T1, T1ce, T2, FLAIR) were provided to the model as input.

3.3. Network Architecture
Due to its local and global feature extraction technique, U-Net is a prominent convolutional neural network used for biomedical semantic image segmentation. It is made up of two paths: an encoder path that extracts features from the input image and a decoder path that combines these extracted features to form the output mask. It employs skip connections between each level of encoders and their associated decoders, allowing the classifier to consider both low-level and high-level features when determining the segmentation mask. However, various constraints impede U-Net training, such as a lack of data augmentation techniques or a restricted number of available layers. It is also challenging for real-time application scenarios where speed is crucial because it needs a substantial amount of computer resources for training. The gradient vanishing problem is a limitation of a U-Net network that can lead to inefficient learning at early layers. Additionally, a large number of parameters present in the model can lead to overfitting if not trained properly. Low-level features are essential for segmentation masks, as they contain information about boundaries, edges, and location of tumorous regions. As the network goes deeper during down-sampling operations, richer higher-level details become available than lower-level ones, but local detail and location data are lost due to convolutional and nonlinearity operations.

This study introduced a dense-encoder and residual-decoder based attention 3D U-Net model for the task of BTS to address the problem of vanishing gradient in the encoder section of the U-Net model during deep network training. We used 3D U-Net as a baseline in our proposed architecture to develop a novel attention 3D U-Net with encoder-decoder blocks to segment brain tumors from MRI images. Our network architecture is depicted in Figure 3. The encoder section gathers context information from the input image, while the decoder section performs accurate segmentation. In Fig. 3, the network encoder is presented on the left, and the network decoder is shown on the right. We fed four modalities 3D MR images (T1, T1Gd, T2 and FLAIR) with sizes of 120 × 120 × 96. Thus, the input data is a 4D patch with a size of 120 × 120 × 96 × 4. The first convolutional block consists of a 7x7 3D convolution followed by a group norm and ReLu; this produces an output feature map with size 120×120×96×24. The feature map is fed to a max pool layer followed by a double convolution layer with a kernel size of 3, which produces an output with a size of 60 × 60 × 48 × 24. Finally, the feature map goes through four dense and transition layers. Because each dense layer in a dense block has access to the feature maps of all dense layers preceding it, it has global information about the entire block, which is useful for accurate segmentation. The contribution of the output feature maps of each dense layer to the global information is controlled by the growth rate of the dense blocks. A dense layer is made up of one 1 × 1 × 1 convolution layer and one 3 × 3 convolution layer. Following each convolution layer, group normalization and the ReLU activation function are utilized. We first perform a 1x1 convolution with 128 filters to reduce the size of the feature maps, and then perform a more expensive 3x3 convolution with a growth rate of 32 feature maps which uses padding to ensure the dimensions remain constant. Concatenation is then used to combine the output from each layer in order to preserve the information from earlier layers. The size of the feature vectors is reduced by using a transition block. As opposed to a pooling layer, Transition Block performs a 1x1 convolution with 32 filters, followed by another 1x1 convolution with a stride of 2, which reduces the size of the volume and the number of feature maps by half. In order to
reduce the information loss associated with pooling, convolution should be performed with a stride of 2. Each convolution is then followed by a group norm and ReLU layers.

A trilinear interpolation layer is used for upsampling, which is subsequently concatenated with low-level features from the corresponding dense block. The output is then passed through a spatial and channel attention block first. Our proposed attention module combines channel, spatial, and skip connection in parallel. However, combining excited features at the same time runs the danger of making feature learning inconsistent. As demonstrated in Fig. 5, we integrate skip connections to reduce network redundancy and sparsity.

Similar to (Islam et al., 2020) by utilizing 3D inter-spatial and inter-channel feature relationships, we introduce 3D attention units to generate 3D spatial and channel attention. We initially carry out a 1x1xC convolution to group all spatial feature correlations into the HxWx1 dimension, after which we carry out average pooling and feed it to the neural network to produce the 3D attention map. We combine skip-connection to reduce the singularity and sparsity brought on by these parallel excitations, improving segmentation prediction and allowing for more generalized learning.

The structure of the residual block is shown in Fig. 4, which is comprised of two 3 x 3 x 3 convolution layers with stride of 1. Group normalization and ReLU activation functions are used after to each convolution layer.

The output of the residual block is the addition of the output of the last convolution layer and the input of the residual block. The shortcut connection is first passed through an additional 1 x 1 x 1 convolution layer with stride of 1. From the top to the bottom of the encoder, the number of feature map increases and the size of feature map reduces progressively. Specifically, Group Norm layer (GN) and activation layer were applied after each convolution layer. The Group Norm aids in creating deeper and wider networks. It first groups the channels and normalize inside each group. It is a trade-off between LayerNorm and InstanceNorm. When batch sizes are small, group normalization outperforms batch normalization. The batch size in our work is one. As a result, we applied group normalization across the entire network. However, when the batch size is significantly large, GN does not scale as well as BN and might not be able to match the performance of BN. The activation layer implements a rectified linear unit (ReLU) layer. The ReLU is employed to optimize the proposed architecture and consequently enhance the performance.

3.4. Losses and Evaluation metrics

The loss and evaluation metrics utilized to evaluate the performance of the proposed model are explored in detail in this section. By minimizing the loss function, the proposed model learns parameters/features from MR images. In the BraTS datasets, about 98 percent of total voxels are non-tumor (Label 0, commonly known as background), with only a few voxels belonging to tumor (Labels 2, 3, 4). We used a BCE-Dice loss function-based region to overcome the class imbalance problem. This loss combines the Dice loss with the standard binary cross-entropy (BCE) loss, which is commonly used in segmentation.
models. Mixing the two approaches allows for some loss variability while benefiting from BCE's stability.

The equation for multi-class BCE by itself will be familiar to anyone who has studied logistic regression:

\[
J(w) = \frac{1}{N} \sum_{n=1}^{N} H(p_n, q_n) = -\frac{1}{N} \sum_{n=1}^{N} [y_n \log \hat{y}_n + (1 - y_n) \log (1 - \hat{y}_n)]
\]

As evaluation metrics, we used the IoU metric, also known as the Jaccard Index, and the Dice coefficient, also known as the Dice-Sørensen coefficient. The Dice score, which calculates the discrepancies between the segmentation result and the ground truth, is a widely used metric for pixel segmentation. The dice score can be calculated as follows for a given predicted and ground truth segmentation \(X\) and \(Y\), where each tumor voxel is labeled as 1, and non-tumor voxels are labeled as 0.

\[
DSC(X, Y) = \frac{2|X \cap Y|}{|X| + |Y|}
\]

The Jaccard metric is similar to the Dice metric in that it calculates the ratio between the overlap of positive instances between two sets by evaluating the performance of pixel segmentation models and their mutual combined values:

\[
J(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}
\]

4. Experimental Results And Analysis

4.1 Implementation details

The Python programming language and PyTorch as its backend were used to implement the suggested model. Due to constrained memory and computational resources, the model was trained for 100 epochs with a batch size of 1. Lion (EvoLved Sign Momentum) optimizer with a learning rate of 0.0001 and weight decay of 0.01 was used to replace Adam. It is a new optimizer discovered by Google Brain that is purportedly to be better than Adam(w), in Pytorch. Each layer of the network was normalized and made more stable by using the activation function ReLU with group normalization. The studies were carried out using the BraTS 2020 benchmark dataset, with 201 images used for training, 34 images for validation, and 134 images for testing purposes. The detailed information about the hyperparameter values that were provided during the training of the model is given in Table 1.

Table 1: List of hyperparameters used for 3D Attention UNet
3D Attention U-Net clearly outperforms other approaches when it comes to model complexity. It took around 11.8 hours to train the proposed model through 100 epochs, with an execution time of 7.12 minutes per epoch. It can be deployed in a real-world clinical setting because our suggested model's online test time for one person was less than one second. By leveraging more effective computational resources, the model's efficiency may be improved even more. Detailed information about the hardware requirement and computational efficiency of the proposed model is given in Table 2.

Table 2: Hardware requirement and computational efficiency of the proposed model

<table>
<thead>
<tr>
<th>Hyperparameters</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input Size</td>
<td>120 x 120 x 96 x 4</td>
</tr>
<tr>
<td>Learning rate</td>
<td>1e-4</td>
</tr>
<tr>
<td>Weight decay</td>
<td>1e-2</td>
</tr>
<tr>
<td>Batch size</td>
<td>1</td>
</tr>
<tr>
<td>Hidden Layer Activation</td>
<td>ReLu</td>
</tr>
<tr>
<td>Optimizer</td>
<td>Lion</td>
</tr>
<tr>
<td>Loss Function</td>
<td>BCEDiceLoss</td>
</tr>
<tr>
<td>Metrics</td>
<td>Dice and Jaccard</td>
</tr>
<tr>
<td>No. of Epochs</td>
<td>100</td>
</tr>
<tr>
<td>Output size</td>
<td>120 x 120 x 96 x 3</td>
</tr>
</tbody>
</table>

4.2 Results and discussion

Figure 6 shows the bar plot of dice and jaccard score obtained by our network on BraTS2020 validation dataset. Table 3 lists the value of dice and jaccard of ET, WT, and TC in the three regions of our segmentation results. The average value, standard deviation, median value, 25 percent quantile value, and 75 percent quantile value are given for each indicator value. It can be seen from Table 3 that the dice mean of WT, TC, and ET are 0.889, 0.866, and 0.828, respectively, and the Jaccard mean of ET, WT, and TC
are 0.718, 0.809, and 0.779, respectively. The WT region has the highest prediction accuracy in our method, whereas the ET region has the lowest. This is due to the fact that the WT region is the largest and the ET region is the smallest. The greater the region, the easier it is to predict. And the smaller the region, the more difficult it is to predict.

Table 3: The performance of our proposed method on BraTS 2020 validation dataset.

<table>
<thead>
<tr>
<th></th>
<th>Dice</th>
<th>Jaccard</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WT</td>
<td>TC</td>
</tr>
<tr>
<td>Mean</td>
<td>0.889</td>
<td>0.866</td>
</tr>
<tr>
<td>StdDev</td>
<td>0.82</td>
<td>0.117</td>
</tr>
<tr>
<td>Median</td>
<td>0.919</td>
<td>0.91</td>
</tr>
<tr>
<td>25quantile</td>
<td>0.878</td>
<td>0.843</td>
</tr>
<tr>
<td>75quantile</td>
<td>0.937</td>
<td>0.938</td>
</tr>
</tbody>
</table>

The performance of our method is compared with the original 3D U-Net method. Figure 7 and Table 4 list the dice and jaccard score obtained by 3D U-Net on the BraTS2020 validation dataset for each class. The proposed technique achieved a higher dice score for each sub-region of the brain tumor than the baseline work because of the use of the dense block with a transition at the decoder part and residual convolutional blocks at the encoder part. The proposed Attention 3D U-Net model significantly improves the segmentation accuracy of the ET class which was the most difficult class to segment.

Table 4: The performance of original 3D U-Net on BraTS 2020 validation dataset.

<table>
<thead>
<tr>
<th></th>
<th>Dice</th>
<th>Jaccard</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>WT</td>
<td>TC</td>
</tr>
<tr>
<td>Mean</td>
<td>0.885</td>
<td>0.793</td>
</tr>
<tr>
<td>StdDev</td>
<td>0.69</td>
<td>0.169</td>
</tr>
<tr>
<td>Median</td>
<td>0.895</td>
<td>0.849</td>
</tr>
<tr>
<td>25quantile</td>
<td>0.847</td>
<td>0.762</td>
</tr>
<tr>
<td>75quantile</td>
<td>0.94</td>
<td>0.896</td>
</tr>
</tbody>
</table>

Our method was qualitatively compared with original 3D U-Net. We randomly selected a patient and used both our method and original 3D U-Net to segment the brain tumor. Figure 8 shows slices of the four modalities of ground truth of the patient and the slices of the predicted segmentation image using our method and the ground truth.
4.2 Comparison with state-of-the-art studies

In comparison to state of art models for brain tumor semantic segmentation, the proposed method is able to segment these individual tumor regions (WT, TC, and ET) close to the ground truth, as shown in the findings. The most challenging areas to segment were the enhancing tumor and its dispersion with necrosis. When it comes to segmenting augmenting enhancing tumor and tumor core, many existing segmentation models performed badly. The proposed Attention 3D U-Net model, on the other hand, successfully segmented these regions. Table 5 also illustrates the quantitative performance of our model, which performs better than the state-of-the-art techniques for enhancing tumor and tumor core classes. In the case of segmenting the whole tumor, the proposed method achieves good results, comparable to state of art techniques. Overall, the proposed method seems to be a better technique to generate segmented images that are closest to the ground truth.

Table 5: Comparison of proposed 3D Attention UNet model with state-of-the-art methods
<table>
<thead>
<tr>
<th>Reference /Study</th>
<th>Image Dimension</th>
<th>Dataset</th>
<th>Dice Score (DSC)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Tumor Core (TC)</td>
</tr>
<tr>
<td>W. Wang et al. (Wang et al., 2021)</td>
<td>128 x 128 x 128</td>
<td>BraTS 2020</td>
<td>0.8173</td>
</tr>
<tr>
<td>O¨.C¸ i¸cek et al. (Cicek et al., 2016)</td>
<td>128 x 128 x 128</td>
<td>BraTS 2020</td>
<td>0.7906</td>
</tr>
<tr>
<td>J. Colman et al. (Colman et al., 2021)</td>
<td>2DSlices with 240 x 240</td>
<td>BraTS 2020</td>
<td>0.7983</td>
</tr>
<tr>
<td>H. Messaoudi et al. (Messaoudi et al., 2021)</td>
<td>192 x 160 x 108</td>
<td>BraTS 2020</td>
<td>0.7520</td>
</tr>
<tr>
<td>L. M. Ballestar et al. (Ballestar and Vilaplana, 2020)</td>
<td>64 x 64 x 64</td>
<td>BraTS 2020</td>
<td>0.7526</td>
</tr>
<tr>
<td>M. Ghaffari et al. (Ghaffari et al., 2021)</td>
<td>128 x 128 x 128</td>
<td>BraTS 2020</td>
<td>0.82</td>
</tr>
<tr>
<td>F. Wang et al. (Wang et al., 2020)</td>
<td>128 x 128 x 128</td>
<td>BraTS 2019</td>
<td>0.798</td>
</tr>
<tr>
<td>J. Zhang et al. (Zhang et al., 2020)</td>
<td>128 x 128 x 128</td>
<td>BraTS 2019</td>
<td>0.777</td>
</tr>
<tr>
<td>A. Myronenko (Myronenko, 2018)</td>
<td>160 x 192 x 128</td>
<td>BraTS 2018</td>
<td>0.8154</td>
</tr>
<tr>
<td>Yang et al. (Yang and Yang, 2018)</td>
<td>96 x 96 x 96</td>
<td>BraTS 2018</td>
<td>0.789</td>
</tr>
<tr>
<td>Raza et al. (Raza et al., 2023)</td>
<td>128 x 128 x 128</td>
<td>BraTS 2020</td>
<td>0.8357</td>
</tr>
<tr>
<td>Proposed model</td>
<td>120 x 120 x 96</td>
<td>BraTS 2020</td>
<td><strong>0.866</strong></td>
</tr>
</tbody>
</table>

Table 6 shows the training time, prediction time, and number of parameters. With 29.5 MB million parameters, our model takes about 11.8 hours to train and 52.65 seconds to predict. This demonstrates that our model is substantially lower in size and trains faster while maintaining the same performance. Due to their computationally efficient designs, new transformer-based architectures improved inference.
times relative to other architectures, however they still fall short of our suggested model in terms of performance.

Table 6: Complexity of models used in comparison process.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Training Time (h )</th>
<th>Prediction Time (s )</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Our method</td>
<td>11.8 hours</td>
<td>52.65</td>
<td>29.5 mln</td>
</tr>
<tr>
<td>3DU -Net</td>
<td>15.2</td>
<td>76.2</td>
<td>19 mln</td>
</tr>
<tr>
<td>3DRes -Net</td>
<td>24.4</td>
<td>65.1</td>
<td>46.4 mln</td>
</tr>
<tr>
<td>Swin BTS</td>
<td>13.2</td>
<td>58.7</td>
<td>16.3 mln</td>
</tr>
<tr>
<td>VT U -Net</td>
<td>15.8</td>
<td>59.8</td>
<td>18.4 mln</td>
</tr>
<tr>
<td>3DLI -Net ++</td>
<td>25.6</td>
<td>64.5</td>
<td>51.6 mln</td>
</tr>
<tr>
<td>Attention 3DUI -Net</td>
<td>18.3</td>
<td>121.4</td>
<td>22.3 mln</td>
</tr>
<tr>
<td>3DRes -U -Net</td>
<td>29.2</td>
<td>69.3</td>
<td>58.6 mln</td>
</tr>
<tr>
<td>3DDeepLab V3</td>
<td>35.8</td>
<td>78.6</td>
<td>76.1 mln</td>
</tr>
<tr>
<td>3DPSP -Net</td>
<td>36.5</td>
<td>96.9</td>
<td>72.4 mln</td>
</tr>
</tbody>
</table>

5. Conclusion

By combining the advantages of ResNet and DenseNet, this paper proposes a deep 3D Attention UNet with dense encoder blocks and residual decoder blocks to increase the performance of brain tumor segmentation from MRI images. The use of dense blocks has the advantage of reducing network parameters, deepening network layers, strengthening feature propagation, avoiding the vanishing-gradient problem, and enlarging receptive fields. In comparison to typical convolution blocks, residual blocks address the issue of network performance degradation as network depth increases. We also introduced 3D attention units to produce 3D spatial and channel attention by leveraging 3D inter-spatial and inter-channel feature correlations. The proposed method surpassed state-of-the-art models, yielding dice scores of 0.866, 0.889, and 0.828 for TC, WT, and ET, respectively, without the use of substantial post-processing. Furthermore, our model takes approximately 11.8 hours to train and 52.65 seconds to predict with only 29.5 MB million parameters. This shows that while maintaining the same performance, our model is significantly smaller and trains faster than other models. By using this approach in clinical oncology, radiologists and oncologists can better understand the different tumor locations, sizes, and shapes, which will aid them with tumor diagnosis, treatment planning, and prognosis. More 3D-based architectures should be investigated in order to minimize computational costs while maximizing contextual information. Post-processing approaches can be
utilized to improve the performance of the model, and this research can be applied to clinically challenging medical imaging problems and other segmentation applications.

**Declarations**

**Acknowledgement**

We want to thank Northwestern Polytechni cal University for helping us to conduct research in this area and providing resource.

**Competing interests:** The authors declare no competing interests.

**References**


---

**Figures**
Figure 1

Four MRI modalities axial slices of a patient data from the BraTS2020 training dataset. From (a) to (d) are T1, T1Gd, T2 and FLAIR. (e) The ground truth segmentation, that purple indicates peritumoral edema, yellow indicates non enhancing tumor core, blue indicates GD-enhancing tumor
Figure 2

Preparing the data and pre-processing workflow
Figure 3

Network Architecture
Figure 4

Building Blocks Explanation

Figure 5

Visual representation of the 3D spatial and channel attention with skip connection.
Figure 6

Dice and Jaccard Coefficients Mean from Validation of our model.

Figure 7

Dice and Jaccard Coefficients Mean from Validation of 3D U-Net.
Figure 8

Qualitative segmentation results of our model on the BRATS-2020 dataset.