BRAIN TUMOR SEGMENTATION USING U-NET ENHANCED WITH ATTENTION

by

SHUYI LI

A thesis submitted to the
Department of Computer Science
in conformity with the requirements for
the degree of Master of Science

Bishop’s University
Canada
December 2023

Copyright © Shuyi Li, 2023
Abstract

Medical image analysis and processing are critical components of clinical applications and scientific research. Deep learning has revolutionized medical image analysis by enabling the identification of morphological or texture patterns from data. In order to enhance brain tumor image analysis, an improved U-Net based model called ArUnet is proposed in this thesis. By integrating convolutional networks and self-attention mechanisms, this thesis focuses on optimizing the ability to extract semantic information in the U-Net network structure, and assists with Residual Net to reduce the problem of difficulty in fitting during the training process. This increases the depth of the model and strengthens U-Net’s own ability to extract abstract features. In addition, self-attention mechanisms help the decoder focus on crucial information, thereby enhancing the model’s performance. The model achieves state-of-the-art performance on the BraTS2021 dataset, with up to 95.54% accuracy.

**Keywords:** Medical Image Analysis, Brain Tumor Segmentation, Artificial Intelligence, Deep Learning, U-Net, Self-attention.
Acknowledgements

I’d like to take a moment to express my gratitude to everyone who has supported me throughout my academic journey. First and foremost, I want to thank my parents for their love and encouragement from my hometown across the ocean. I miss them dearly, and their love has always been the driving force behind my efforts, motivating me to live up to their expectations. I also want to extend my thanks to my supervisor, professor Madjid Allili and professor Mohammed Ayoub Alaoui Mhamdi. Their patient guidance and encouragement have filled my academic journey in a foreign land with warmth and confidence. Most importantly, I want to express my appreciation to my boyfriend for his guidance and assistance, allowing me to fearlessly pursue my dreams in a life filled with love. Additionally, I am grateful to Bishop’s University, especially the Computer Science Department, for granting me the opportunity to come to Canada and engage in research and learning in a field I’m passionate about. Likewise, I want to thank the Graduate Entrance Scholarship Committee of Bishop’s University and the MITACS Foundation for their financial support. Their help has significantly eased the pressures in my life, enabling me to better concentrate on my studies. Lastly, I’d like to express my gratitude to my two hometown cats, Aimi and Aiwa. In times of difficulty and low spirits, their adorable presence has offered solace to my soul. Thank you all!
# Contents

## 1 Introduction

1.1 Why Computer-aided Brain Tumor Diagnosis? ..................................................... 1  
1.1.1 The Necessary Steps of Brain Tumor Diagnosis .............................................. 2  
1.1.2 Brain Tumor Scanning Technology and Medical Imaging .............................. 3  
1.2 Applying Deep Learning Methods in Brain Tumor Segmentation Images .......... 5  
1.2.1 Research on Automatic Segmentation Systems .............................................. 6  
1.3 The Advantages of U-Net in Realizing Tumor Segmentation ............................ 8  
1.3.1 Encoder and Decoder Architecture ............................................................... 8  
1.3.2 Achievements Based on U-Net ........................................................................ 9  
1.3.3 The Aspects of U-Net Still Need Improvements ............................................ 10  
1.4 New Possibilities Brought by Attention Mechanism .......................................... 11  
1.4.1 How to Focus Computer’s Attention ............................................................... 11

## 2 Related Work

2.1 Different improvement methods of U-Net ......................................................... 15  
2.2 Advantages of residual networks ..................................................................... 17  
2.2.1 Model Degeneration problem and Residual Networks ................................ 18  
2.2.2 Application of residual block ..................................................................... 19  
2.3 Attention mechanism ...................................................................................... 20  
2.3.1 The combination of attention mechanism and U-Net ................................. 20

## 3 Methodology

3.1 Overall Structure of ArUnet ............................................................................. 22  
3.1.1 U-shape Architecture ................................................................................. 23  
3.2 Residual Modules for Down-sampling and Up-sampling ............................... 24  
3.2.1 The function of residual blocks ................................................................... 24  
3.2.2 Residual connection .................................................................................... 25  
3.3 Feature Fusion via Self-Attention ................................................................. 26  
3.3.1 Matrix Q, K, V of Self-attention ................................................................. 27  
3.3.2 Computing The Attention in Convolutional Neural Networks .................. 28

## 4 Experiment and Results

4.1 Dataset and Data Processing ....................................................................... 30  
4.2 Training Procedure and Evaluation Metrics ................................................... 32  
4.3 Implementation ............................................................................................... 32  
4.4 Results and Analysis ...................................................................................... 34
5 Conclusion and future works

Bibliography
Chapter 1

Introduction

1.1 Why Computer-aided Brain Tumor Diagnosis?

Early diagnosis of brain tumors plays an important role in improving treatment outcomes and patient survival. Brain tumor is considered one of the deadliest diseases in the world due to their increasing impact and mortality in all age groups [1]. Therefore, in order to alleviate the impact of brain tumors on human brain health and lifespan, we must first understand what brain tumors are and the types of brain tumors. Brain tumors are a group of abnormal cells that grow in or around the brain and are generally classified into meningiomas, gliomas, and pituitary tumors [2].

The brain is an important organ in the human body responsible for control and decision-making. As the management center of the nervous system, this part must be protected from any injury and disease. Because of the various shapes, sizes, locations, and appearances of brain tumors in the brain, as shown in Figure 1.1, tumor detection is more difficult [3]. Among gliomas, high-grade glioma (HGG) is considered to be the most aggressive brain tumor type, with an overall survival rate of no more than 14 months [4]. Research from the National Brain Tumor Foundation (NBTF) estimates that 29,000 people in the United States are diagnosed with brain tumors each year, and nearly 13,000 die [5].
CHAPTER 1. INTRODUCTION

Accurate and timely detection of brain tumors is an important task for disease prevention and treatment, especially for tumor cutting surgery. Manual diagnosis by radiologists and other experts takes time to segment and classify brain tumors, and there are significant differences in the diagnosis of brain tumors by different medical experts. Therefore, this thesis aims to design a deep learning network with high accuracy for segmenting brain tumors, in order to provide assistance for the analysis of tumor regions [7] through computer image processing.

1.1.1 The Necessary Steps of Brain Tumor Diagnosis

The diagnosis of brain tumors can be divided into the process of tumor detection, segmentation, and classification [8].

Tumor detection is the process of detecting the presence or absence of tumors using medical tumor image dataset. This is a common issue in the medical field, and the output of the tumor detection process is typically an image of the tumor, labeled as either normal or abnormal. However, the accuracy of this process is heavily reliant on the quality of the imaging technology used. Thankfully, the field of tumor detection has undergone significant development and has reached a mature stage, providing a robust foundation for accurate and reliable tumor diagnosis.

Tumor segmentation is the process of dividing an image into regions of interest to facilitate data characterization, description, and visualization. The purpose of segmentation is generally to enhance the performance of medical images by making the features on the location and boundary of the tumor more important, obvious, and easy to analyze. In practical operations, several methods
and techniques are used for the segmentation of brain tumors, including artificial neural network segmentation techniques [9], edge-based algorithms [10], and clustering algorithms [11]. In this field, U-Net is a model with an encoder-decoder structure, combined with skip connections before the decoder, so that the encoder can access high-resolution features layer by layer during the upsampling process. Hence, this thesis will establish a deep learning image segmentation framework for brain tumor glioma MRI images based on the combination of this encoder-decoder structure.

Tumor classification is the process of assigning input features to different categories or classes. Feature extraction and selection are critical for classification, particularly in the case of brain tumors. The primary objective of brain tumor classification is to categorize tumors as either benign or malignant or to grade them based on MRI images. The accuracy of this process depends on the quality of brain tumor segmentation, with higher completeness of segmentation details leading to greater accuracy in classification.

To summarize, compared to manual methods, using computer-assisted methods for detecting, segmenting, and classifying brain tumors can greatly improve accuracy and speed while reducing noise. This is a crucial premise of the method proposed in this thesis, as it enables the utilization of traditional machine learning and deep learning approaches to enhance computer-assisted tumor diagnosis, particularly in the area of tumor segmentation. However, the quality of computer-assisted diagnosis is contingent on the quantity and quality of the database, which is determined by medical imaging technology and expert labeling. In the following section, we will concentrate on the impact of medical imaging technology on the proposed method.

1.1.2 Brain Tumor Scanning Technology and Medical Imaging

Early detection of brain tumors can help radiologists make effective prognoses and increase the chances of long-term survival. Medical imaging is an important diagnostic tool that can greatly assist radiologists in image analysis and diagnosis. Over the years, various techniques have been used to detect brain abnormalities, including computed tomography (CT), positron emission tomography (PET) scans, and magnetic resonance imaging (MRI), as shown in Figure 1.2. CT, uses X-ray scanning to obtain internal structural images of the body. It is commonly used for evaluating the skeletal system and organs, providing high anatomical details, but with lower resolution in certain soft tissues. PET, on the other hand, utilizes radioactive isotopes labeled drugs to observe the metabolic activity and functional status of body tissues. It is suitable for cancer diagnosis, cardiovascular disease assessment, and brain function research, but it lacks clear anatomical details. MRI employs a strong magnetic field and radio waves to generate detailed anatomical structures and tissue information. It offers higher resolution in examinations of the brain, spine, joints, and soft tissues, with an advantage in imaging soft tissues. Recently, computer-aided analysis of medical images has been implemented in clinical settings to assist radiologists in image interpretation and detect abnormalities. For example, CT imaging can help doctors identify tumors and accurately determine their size and location [12]. Advanced imaging and computing technologies are facilitating work in image-guided procedures and more realistic visualizations.
CHAPTER 1. INTRODUCTION

Figure 1.2: Medical images produced by different scanning methods.

The emergence of magnetic resonance imaging (MRI) has opened up new avenues for human exploration. MRI can capture differences between malignant and benign tissue, and after decades of development, the latest MRI technology can generate extremely accurate internal anatomical images.

MRI is the most commonly used and effective technique for detecting brain tumors. As a non-invasive in vivo imaging technique, MRI stimulates the target tissue with radiofrequency signals under the influence of a very strong magnetic field to produce internal images. In current routine clinical examinations, the four standard MRI modes for diagnosing gliomas include T1-weighted MRI (T1), T2-weighted MRI (T2), T1-weighted MRI gadolinium-enhanced (T1-Gd), and fluid-attenuated inversion recovery (FLAIR), as shown in Figure 1.3. Currently, MRI is gradually being applied to computer-aided diagnosis, which can obtain better accuracy and efficiency than manual methods. Current research has begun to focus on computer-based tumor detection and segmentation using various machine learning and deep learning techniques [13]. However, considering the heterogeneity of tumors in terms of visual features, spatial distribution, morphology, and size, further use of various machine learning and deep learning techniques to improve computer-aided diagnosis is still a challenging task. This thesis aims to contribute to this field by using MRI images, which are currently more widely used, as the research object.
CHAPTER 1. INTRODUCTION

1.2 Applying Deep Learning Methods in Brain Tumor Segmentation Images

The medical image segmentation technique discussed in this thesis is fundamentally important for analyzing brain tumor medical images. Image segmentation is to divide an image into meaningful and visually coherent regions or objects based on certain features or attributes. The purpose of medical image segmentation is to divide the image into different regions with similar attributes, including grayscale intensity, texture, color, contrast, boundaries, and brightness. Segmentation is the primary task of MRI processing, which provides us with a means to understand tumor images.

However, the brain tumor MRI image segmentation task discussed in this thesis has always faced enormous challenges. Firstly, brain tumor images generally contain multiple intertwined and overlapping tissues, including tumor tissue (solid tumors, edema, necrosis) and normal brain tissue, such as gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF), as shown in Figure 1.4. The brain tumor studied in our thesis is a glioblastoma, an invasive tumor with blurred boundaries that are often difficult to distinguish from healthy tissue [15]. Therefore, the complex and diverse morphological and structural characteristics of brain tumors provide opportunities for further exploration in the field of computer vision. Additionally, the significant variability in the size, spatial range, and location of brain tumor structures hinders the applicability of machine learning segmentation algorithms that rely on strong prior knowledge [9].
Currently, MRI sequences are the most commonly used imaging method for tumor compartmentalization in clinical practice. MRI images have better clarity and contrast and are more suitable for various deep learning segmentation methods than other imaging methods, such as CT [17]. Most current research focuses on using computer algorithms for automatic tumor segmentation, which may provide objective, reproducible, and scalable methods for the quantitative evaluation of brain tumors. The following content will introduce the application of deep learning in the field of tumor image segmentation.

1.2.1 Research on Automatic Segmentation Systems

The automatic segmentation system involves many methods, and we are committed to using a supervised learning model of deep learning to accomplish the automatic segmentation of brain tumors. Automatic segmentation methods offer efficient and objective segmentation, which means that images can be automatically and objectively partitioned into meaningful and coherent regions or objects without relying on subjective input or manual intervention. It has become an interesting and popular research field in recent years [18], so many researchers have developed various brain tumor segmentation methods, including fuzzy clustering method, machine learning method, and deep learning method. Classification methods such as k-nearest neighbors, support vector machines (SVM), and naive Bayes are popular machine learning techniques in image analysis in recent years. Applying statistical learning methods to low-level brain tumor classification features is common in traditional machine learning methods, which mainly focus on the estimation of tumor boundaries and their localization. Furthermore, they rely heavily on preprocessing techniques for contrast enhancement, image sharpening, and edge detection/refinement, thereby relying on human expertise for feature engineering [19]. However, feature engineering relies on manual selection and identification of features, often struggling to capture complex patterns and relationships in the data, and it also incurs significant time and computational costs. Since deep learning automatically learns features from raw data and can discover new features that are relevant to specific applications [9], it has a higher capacity to capture nonlinear relationships, hierarchical structures, and interactions between features. Deep learning-based approaches have been shown to outperform machine learning
methods in the processing and analysis of medical images.

Deep learning, a subfield of machine learning that relies on the availability of large-scale datasets for training and requires smaller preprocessing steps than traditional methods, has dominated the field of brain tumor segmentation over the past few years. Deep learning approaches address representation learning by automatically learning increasingly complex feature hierarchies directly from the data [20]. For example, convolutional neural networks (CNNs) have powerful capabilities in constructing hierarchical task-specific feature representations through end-to-end training networks, Havaei et al. [9] have successfully integrated convolutional neural networks (CNNs) into deep neural network-based segmentation frameworks, used to segment brain tumors in MRI images. Different from the input of hand-crafted features in traditional classification methods, the main advantage of CNNs is the ability to recognize the most important patterns and information from training images [21]. CNNs take input patches extracted from images and use trainable convolutional filters and local subsampling to extract increasingly complex feature hierarchies. Due to this characteristic, research on CNN-based brain tumor segmentation mainly focuses on network structure design.

Several deep learning architectures have demonstrated success in brain tumor segmentation, including VGGNet [22], GoogleNet [23], and AlexNet [24], which have been successful for image classification. Zikic et al. [25] designed a shallow Convolutional Neural Network (CNN) with two convolutional layers separated by max pooling, followed by a fully connected layer and a softmax layer for segmentation experiments on brain tumor tissue. Lyksborg et al. [26] utilized an ensemble of 2D CNNs, as showed in Figure 1.5, combined with a grow-and-cut approach to identify intact tumors, which were then segmented smoothly by cellular automata and refined by multi-class CNNs. Rao et al. [27] extracted patches in each plane of each voxel and trained a separate CNN for each MRI sequence; the outputs of the last fully connected layer were concatenated and used to train a random forest classifier. These methods have achieved impressive performance on challenging datasets, demonstrating the effectiveness of CNNs in tumor segmentation from medical scans.

Figure 1.5: The network structure proposed by Lyksborg [26].
In 2014, Long et al. [28] proposed a new end-to-end fully convolutional network (FCN), which revitalized the field of natural image segmentation and was applied to brain tumor segmentation. In 2014, Ross et al. [29] introduced R-CNN (Regions with CNN features), which combines the utilization of high-capacity convolutional neural networks (CNN) with bottom-up region proposals for object localization and segmentation. This laid the groundwork for Mask R-CNN, proposed by Kaiming et al. [30] in 2017, which added an additional segmentation branch to R-CNN, enabling pixel-level instance segmentation. However, building upon the FCN model, U-Net, introduced by Ronneberger et al. [31] in 2015, holds greater significance in the field of biomedical image segmentation tasks. The next section will focus on U-Net’s relevant research and elaborate on its significance in brain tumor segmentation, laying the foundation for this thesis’s study.

1.3 The Advantages of U-Net in Realizing Tumor Segmentation

Among various CNN variants, U-Net, a typical codec-based network, exhibits excellent segmentation potential and becomes the mainstream architecture for medical image segmentation.

1.3.1 Encoder and Decoder Architecture

U-net is a neural network structure mainly used for image segmentation [31]. As shown in Figure 1.6, this thesis mainly utilizes the two basic paths provided by the U-net architecture. The first path is the contraction path, also known as the encoder or analysis path, which is similar to a regular convolutional network and provides classification information. The second is the expansion path, also known as the decoder or synthesis path, which consists of up-convolution and concatenation with features from the contraction path. This extension allows the network to learn localized taxonomic information and restore the image contours. We can use this expansion path to increase the output resolution, which can then be passed to the final convolutional layer to create a fully segmented brain tumor image.

![Figure 1.6: Structure of U-Net [31].](image-url)
The U-Net architecture’s contraction path follows the standard design of convolutional neural networks, the original author employing two 3x3 convolutions without padding and a ReLU activation function, followed by a 2x2 max pooling operation to enable downsampling. At each downsampling stage, the number of feature channels doubles. ReLU (Rectified Linear Unit) activation function helps avoid the vanishing gradient problem by outputting zero for negative inputs and the input value itself for positive inputs, without saturating in the positive range, thus preventing gradients from becoming very small. In the field of medical imaging, unlike the sigmoid activation function that saturates at extreme input values, ReLU is unaffected by the vanishing gradient problem, promoting faster and more stable learning in deep networks.

The expansion path features a 2x2 up convolution to upsample the feature map, reducing the number of feature channels by half. The upsampled map is then concatenated with the corresponding feature map from the contraction path, cropped to match the resolution, and processed by two 3x3 convolutions with ReLU activation. Clipping is necessary to address the loss of boundary pixels that occur in each convolution. In the final layer, a 1x1 convolution is applied to map each 64-component feature vector to the desired number of classes. The U-Net architecture consists of 23 convolutional layers. To ensure a seamless tiling of the output segmentation map, the input tile size must be selected such that all 2x2 max pooling operations are performed on layers of dimensions $x$ and $y$.

We can visualize the U-Net architecture as a symmetrical U-shaped structure. Ronneberger [31] also employ a symmetrical encoder-decoder structure with skip connections to enhance detail preservation. The main idea is to downsample the fixed-sized image using the encoder to adapt to the receptive field’s size and extract deep features of the image. Then, the downsampled results are fed into the decoder in a cascaded manner for upsampling to restore the image’s size and restore the spatial relationship between pixels by merging the intermediate features generated at each downsampling stage of the encoder through skip connections. Finally, convolutional layers can learn to assemble a more accurate output from this information. The U-Net network provides us with an important advantage, which is a large number of feature channels in the upsampling part, allowing our model to propagate contextual information to higher resolution layers [32]. In addition, we utilize the feature map generated by U-Net, which combines both low-level and high-level features while preserving positional information, to fuse features from different levels and restore more accurate segmentation edges, greatly improving the segmentation performance of brain tumor images.

1.3.2 Achievements Based on U-Net

The U-net network was developed based on the work of Long [28], using a fully convolutional network. In the ISBI cell tracking challenge in 2015, Olaf Ronneberger et al. proposed the concept of a U-net network [31] to apply end-to-end training to medical image analysis. Currently, U-Net has become a benchmark in solving brain tumor segmentation tasks. At the same time, various improved U-Net methods have been proposed, such as UNet++ [33], Residual U-Net [34], Inception U-Net [35], UNETR [36], Dense U-Net [37], R2U-Net [38], KiU-Net [39], Attention U-Net [40], and U-Net 3+ [41], as shown in Table 1.1. These modifications aim to enhance the accuracy and efficiency of inter-layer and intra-layer feature transfer in the U-Net network. For instance, Dense U-Net [37] combines the concepts of densely connected networks and U-Net for 3D liver and tumor segmentation. UNet++ [33] improves upon U-Net by using dense skip connections between different layers in the encoder and decoder. Meanwhile, MDU-net [42] introduces multi-scale dense connections in the U-Net architecture. Finally, FU-net [43] enhances U-Net by proposing a dynamically weighted cross-entropy loss function. These advancements in U-Net-based architectures have demonstrated that well-designed U-Net networks can effectively perform tasks such as brain tumor
CHAPTER 1. INTRODUCTION

segmentation. This provides ample support for our research and will also be the focus of this thesis, utilizing residual networks or attention mechanisms to improve the feature extraction efficiency of the hidden layers in U-Net, with the overall goal of enhancing both accuracy and training efficiency.

<table>
<thead>
<tr>
<th>Network</th>
<th>Architecture</th>
<th>Advantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>U-Net [28]</td>
<td>U-shape architecture; Skip connection</td>
<td>High-resolution prediction; Effective feature reuse</td>
</tr>
<tr>
<td>U-Net++ [33]</td>
<td>Multi-scale feature fusion; Pyramid pooling</td>
<td>Increased depth; Multi-resolution prediction</td>
</tr>
<tr>
<td>U-Net3+ [41]</td>
<td>Parallel paths; Residual blocks</td>
<td>Efficient training; Improved accuracy</td>
</tr>
<tr>
<td>Res-UNet [34]</td>
<td>Residual connections</td>
<td>Preserve multi-level features</td>
</tr>
<tr>
<td>Inc-UNet [35]</td>
<td>Inception modules</td>
<td>Combined features of multiple types</td>
</tr>
<tr>
<td>U-Net-Transformer [36]</td>
<td>Self-attention mechanism</td>
<td>Better predictions</td>
</tr>
<tr>
<td>Dense-UNet [37]</td>
<td>Dense connections</td>
<td>All layers connected for more complex features</td>
</tr>
<tr>
<td>R2-UNet [38]</td>
<td>Residual connections and recurrent connections; Batch normalization and Dropout regularization</td>
<td>Improved accuracy</td>
</tr>
<tr>
<td>Ki-UNet [39]</td>
<td>Implemented using Keras framework</td>
<td>Basically equals to Unet</td>
</tr>
<tr>
<td>Att-Unet [40]</td>
<td>Attention mechanism; Deeper layers</td>
<td>Focus on import features; Computationally expensive</td>
</tr>
</tbody>
</table>

Table 1.1: Summary of different U-net improvements.

1.3.3 The Aspects of U-Net Still Need Improvements

Although previous studies using the U-Net method have achieved significant success, they are limited in their ability to make substantial progress. This is due to the characteristics of the U-Net architecture, which decodes the feature maps we want through high resolution features and then passes them through upsampling, creating a built-in bias. Furthermore, in this convolutional network, each kernel in each convolutional layer can only focus on a sub-region of the entire image, causing a loss of global context and the ability to establish long-term dependencies. Additionally, the dimension and configuration of convolutional filters are usually fixed, resulting in a loss of information when using skip connections and affecting the final prediction [44]. To tackle challenges like BraTS21, selecting the best neural network architecture and training plan simultaneously is necessary but a daunting task for improving the U-Net architecture based on previous studies. To address these issues, we choose to combine the attention mechanism from natural language processing (NLP) with the U-Net model to overcome the inherent drawbacks in modeling long-range dependencies.
1.4 New Possibilities Brought by Attention Mechanism

It is well known that the attention mechanism is an effective tool for enhancing feature extraction in medical image segmentation. It allows for the modeling of long-term relationships between feature representations and updated feature maps, making it possible to better capture the crucial long-range dependencies required for the accurate segmentation of organs and tumors. In recent years, the integration of attention mechanisms in medical image segmentation has gained widespread attention and has achieved impressive results [45]. Therefore, this thesis intends to improve the image segmentation performance of U-Net using self-attention mechanism.

We know that attention mechanisms originated from human visual cognitive science, and attention mechanism plays a vital role in human perception [46]. The fundamental characteristic of perception is that it involves selective processing, as humans do not tend to process complete information all at once. Instead, individuals allocate their attentional resources to salient information, while disregarding irrelevant sensory input. For instance, during visual perception, humans do not attend to the entire scene in a linear fashion, but rather selectively focus on relevant features as needed. Upon repeated exposure to a scene containing high-value information in a particular region, individuals learn to selectively attend to that area, thereby enhancing their ability to efficiently extract pertinent information from complex stimuli with limited cognitive resources. Using this adaptive mechanism to improve deep learning models will bring new perceptions to machines and enhance their ability to perceive and process human knowledge. Next, we will mainly introduce the relevant content of self-attention mechanism.

1.4.1 How to Focus Computer’s Attention

![Figure 1.7: Structure of transformer network [47].](image-url)
In transformer networks, as shown in Figure 1.7, the attention mechanism is utilized to process input data which is represented as a collection of queries (Q), keys (K), and values (V), as shown in Figure 1.8. The queries represent the current input representation for the network while the keys represent the contextual information of the input. The attention scores are computed by taking the dot product of the queries and keys, which measures the similarity between them. The attention scores determine the importance of each token in the input sequence, and a weighted sum of the values (V) is calculated using these scores. The weighted sum reflects the relative importance of each token as determined by the attention scores. The self-attention mechanism produces an output in the form of the weighted sum of the values which is then passed on to the next layer in the network. This thesis will utilize such dynamic mechanisms to enable the network to selectively focus on different parts of the input sequence based on the relationships between the tags in the input, thereby enhancing the network’s ability to capture contextually relevant relationships.

In natural language processing (NLP) tasks, the ordering of tokens in a sequence is often of considerable importance, and disregarding this ordering can lead to inaccurate predictions. In our project, we chose to address this issue by incorporating positional embeddings into the input data to provide information about the relative position of each token in the sequence. These embeddings are learned representations of the relative position of each token and their inclusion in the input data before being processed by the network enables the network to consider the order of the tokens, as shown in Figure 1.9. This is achieved by permitting the network to differentiate between the various positions of the tokens in the input sequence, thus allowing it to better capture the relationships between tokens in the input. Therefore, the combination of positional embeddings is crucial for ensuring the accuracy of the predictions made by the self-attention mechanism in transformer networks for NLP tasks. In the image segmentation task studied in this thesis, the ordered sequence of tokens plays a very important and positive role in preserving the spatial characteristics of the image.
Incorporating attention mechanisms in deep learning networks can greatly improve the efficiency and accuracy of perceptual information processing. This is because attention mechanisms can dynamically allocate input weights of neurons to selectively focus on the most crucial parts of information. Therefore, this thesis will focus on incorporating the attention mechanism into the encoder-decoder structure based on U-Net to accomplish the task of extracting and restoring near-boundary features in brain glioma images, thereby enhancing the image segmentation capability of CNN models. Numerous studies have shown that by adding attention modules, deep CNNs can accelerate the learning process, extract more critical and discriminative features for target tasks, enhance the robustness of the network model, and better adapt to smaller training datasets. The attention mechanism was originally proposed by the Google DeepMind team when performing image classification tasks, thus starting a wave of research on attention mechanisms [50]. Attention mechanisms first appeared in natural language processing (NLP) and quickly gained dominance. Wang et al. [51] proposed to combine non-local blocks with existing architectures, and introduced the attention mechanism into computer vision for the first time. The attention mechanism is currently used in machine translation [52], speech recognition [53], computer vision [40], and many other fields have a wide range of applications.

Attention mechanisms are effective in capturing long-term dependencies and important responses in the field of computer vision. Wang et al. [51] proposed non-local neural networks (NL-Nets), which capture long-term dependencies by aggregating query-specific global context to each query location. This technique has demonstrated promising results in various applications, including image recognition and video classification. Meanwhile, Hu et al. proposed Squeeze and Excite Networks (SE-Nets) [54] to process global spatial information of various channels in a soft attention manner, i.e., learning and rescaling scale factors. In addition, the attention mechanism has also been successfully introduced into the field of medical image segmentation [55]. Kaul et al. [56] propose a method for incorporating attention into an FCN, FocusNet, for medical image segmentation from feature maps generated by a separate convolutional autoencoder. This thesis proposes a novel fusion method to apply attention mechanism to image segmentation tasks in deep learning models, aiming to address the significant challenges of tumor image segmentation in the automatic diagnosis field emphasized in the previous text.

Since the U-Net model gradually recovers the downsampled images, the attention mechanism can help connect the information flow from deep layers to shallow layers, improving the accuracy of upsampling. In light of this, the present study aims to integrate the attention mechanism into the
U-Net model for brain tumor segmentation and to explore the influence of local responses on its accuracy.

The remaining chapters of this thesis are arranged as follows. Chapter 2 discusses three objects related to the methodology of this study: U-Net, Residual Networks, and Attention Mechanism. It provides a detailed introduction to their technical backgrounds and research status. Chapter 3 provides a detailed description of the model structure in this study, along with implementation details of two main techniques. This includes how to utilize residual modules for sampling and how to employ self-attention mechanisms for feature fusion. Chapter 4 elaborates on the experiments and results. Chapter 5 concludes the article with a summary and outlines future work.
Chapter 2

Related Work

2.1 Different improvement methods of U-Net

In [57], the authors proposed a 2D fully convolutional complex segmentation network based on the U-Net architecture to accurately segment brain tumors. The network architecture is identical to U-Net, but to improve segmentation accuracy, they employed comprehensive data enhancement techniques, including rigid and affine deformation, brightness transformation, and elastic distortion. The model was evaluated on the well-established BRATS 2015 benchmark dataset, confirming the reliability of U-Net and suggesting a promising direction for future medical image analysis research: improving the performance of U-Net through data augmentation. To improve segmentation accuracy in medical images, a new and more powerful architecture called UNet++ [33] has been proposed. The architecture can be described as a deeply supervised encoder-decoder network, where a series of nested dense skip paths connect the encoder and decoder subnetworks. Unlike U-Net, UNet++ uses dense convolutional blocks to process the encoder feature map before passing it to the decoder, which makes the semantic level of the encoder feature map closer to that of the decoder feature map. This approach reduces the semantic gap between encoder and decoder feature maps, making optimization easier. The redesigned skipping path and deep supervision in the UNet++ architecture improve the accuracy of medical image segmentation, which suggests that optimized skipping paths can be a promising method for medical image analysis. Figure 2.1 illustrates the nested and dense skip connections used in the UNet++ architecture.
In [58], the authors developed a fully automatic and efficient brain tumor segmentation method based on a 2D deep convolutional neural network (DNN) to extract whole tumor and intratumoral regions. The method is quantitatively evaluated on the BraTS2018 training dataset and challenge validation dataset, demonstrating satisfactory segmentation results. This study highlights the universal applicability of U-net to multimodal MRI data and the importance of using an appropriate loss function to optimize segmentation performance. As deep learning tasks continue to evolve, the improved version of U-Net has gradually become insufficient, prompting researchers to incorporate advanced ideas from other fields to address its shortcomings. In [59], the authors proposed USE-Net, an efficient convolutional neural network (CNN) and one of the most effective CNNs in biomedical image segmentation, was proposed. Built on the U-Net architecture, USE-Net incorporates Squeeze and Excite (SE) modules to capture dataset-specific features and improve generalization performance. Unlike the original SE block, USE-Net adds the SE module after each encoder and decoder block, enabling adaptive channel-level feature recalibration. The study demonstrates USE-Net’s ability to analyze multiple heterogeneous MRI datasets with high accuracy and compares its performance with other state-of-the-art CNN-based architectures. The results suggest that attention mechanisms with feature recalibration adaptive mechanisms, similar to SE modules, may be valuable solutions for medical imaging applications.

U-Net is a down convolutional autoencoder that uses a standard codec architecture to reduce data dimensionality near the bottleneck. The encoder reduces the dimensionality of the data, enabling the network to focus on higher-level features with increasing depth. However, U-Net’s ability to capture fine features is limited by the smallest receptive field in the network, which is determined by the first layer. This can pose challenges for incomplete architectures like U-Net in capturing fine details. In recent studies, researchers proposed Ki-Net as a solution to this problem [39]. Ki-Net projects data to a higher dimension, resulting in a larger spatial dimension in the intermediate layer than that of the input data, as shown in Figure 2.2. To achieve this, an up-sampling layer is used after each convolutional layer in the encoder, which enlarges the spatial dimension of the intermediate layer, while a max pooling layer is used after each convolutional layer in the decoder to reduce the dimensionality back to that of the input. Ki-Net exhibits superior edge-capturing ability.
compared to U-Net and improves overall segmentation accuracy when combined with U-Net. The
augmented network, KiU-Net, outperforms U-Net in segmenting small anatomical landmarks and
smoothing out noise boundaries while having faster convergence and fewer parameters. By combi-
ing Ki-Net’s fine edge-capturing feature maps and U-Net’s shape-capturing feature maps, KiU-Net
can leverage the benefits of both architectures, resulting in improved segmentation accuracy. Addi-
tionally, a cross-scale based residual block fusion strategy is introduced, which efficiently utilizes
information from the two networks. KiU-Net achieves state-of-the-art performance on the brain
anatomy segmentation task of US images, indicating that a multi-channel network structure can ex-
tract more complete features, and appropriate feature fusion techniques can significantly enhance
convolutional networks’ performance in medical image segmentation tasks.

Figure 2.2: Structure of KiU-Net network [39].

In conclusion, Medical image segmentation is a critical task for many clinical applications, and
U-Net is among the most popular and widely used convolutional neural networks for this purpose.
Over the years, researchers have proposed numerous modifications to U-Net, such as data augmen-
tation, dense and nested skip connections, deep supervision, and adaptive feature recalibration, all
of which have enhanced its accuracy and generalization ability in various medical imaging appli-
cations. However, further improvements could be made by exploring hybrid models that combine
U-Net with other architectures, or by incorporating novel techniques such as transfer learning, at-
tention mechanisms, domain adaptation, improved modules, and generative adversarial networks.
Thus, future research in medical image segmentation could focus on these approaches to further
enhance the accuracy and efficiency of this critical task.

2.2 Advantages of residual networks

Deep learning has proven to be remarkably effective in various computer vision tasks. However, the
introduction of deeper neural networks has given rise to questions regarding the benefits of adding
more layers to improve learning. One of the major challenges in this regard has been the notorious
problem of vanishing or exploding gradients[60, 61], which has impeded progress in answering
this question [34]. Fortunately, recent advances in techniques such as normalized initialization[61,
62, 63] and intermediate normalization layers [64] have resolved the issue of exploding gradients,
making it possible to use deep neural networks in stochastic gradient descent and backpropagation, ultimately leading to successful convergences [65].

Despite the ability of deeper neural networks to converge, a common challenge they face is the degradation problem, where increasing the network depth initially leads to accuracy saturation followed by rapid degradation. Previous research has revealed that augmenting layers in a sufficiently deep model can actually result in higher training error [66, 67]. This phenomenon is exemplified in Figure 2.3, depicting the training error (left) and test error (right) for 20-layer and 56-layer "normal" networks trained on CIFAR-10.

![Figure 2.3: Model degradation phenomenon [34].](image)

### 2.2.1 Model Degeneration problem and Residual Networks

Residual Networks, commonly referred to as ResNets, were introduced by Kaiming He et al. in 2016 as a deep residual learning framework [34, 68] to address the model degradation problem described earlier. In mathematical statistics, residuals denote the difference between the observed and estimated values. In deep learning, training models with many layers can be time-consuming and limited by the number of layers that can be effectively trained. To circumvent this issue, ResNets reprogram the network layers to act as residual functions of their inputs. The introduction of skip connections in ResNets allows the network to use the residual function $y = F(x, \{W_i\}) + x$ to retain the features of the input, leading to better performance.

To demonstrate the effectiveness of their proposed method, He et al. conducted extensive experiments on ImageNet [69] and found that it successfully solves the degradation problem with two key advantages:

1. Compared to ordinary networks that simply stack layers, deep residual networks are easier to optimize. However, as the depth increases, they may suffer from the issue of increasing training error.
2. Deep residual networks achieve considerable improvements in accuracy by increasing their depth, surpassing those of previous networks.

The observations made by He et al. were also confirmed by experiments on the CIFAR-10 dataset [70], indicating that the optimization challenges and advantages of the deep residual learning framework are not limited to specific datasets. These experiments demonstrated that the use of identity mapping is both sufficient to solve the degenerate problem.

Residual networks (ResNets) have shown impressive performance in various computer vision tasks, and have become a vital tool widely used in the field of deep learning. Recent research has focused on exploring the use of residual connections to design efficient architectures that can address
CHAPTER 2. RELATED WORK

19

the issue of model degradation.

2.2.2 Application of residual block

ResNets have proven to be superior to earlier models in a variety of tasks, such as image classification[71, 72, 73], object detection [74], and semantic segmentation [75], and are gradually replacing VGGNets [76] in computer vision. Their most notable advantage is their ability to efficiently train very deep neural networks, which overcomes the vanishing gradient problem that arises when training such networks [34]. Szegedy et al. [77] demonstrated that training with residual connections considerably accelerates the training of inception networks. They also proposed simplified architectures for both residual and non-residual inception networks, which significantly improved the single-frame recognition performance on the ILSVRC 2012 classification task. Additionally, Mahajan et al. [71] extracted image features from a pre-trained ResNet and utilized these features to train a Support Vector Machine (SVM) classifier, showcasing that deep networks make feature extraction easier and faster compared to other traditional network methods.

Deep residual networks are known to improve performance but at the cost of requiring nearly double the number of layers, resulting in reduced feature reuse and slower training times for large networks. To address these issues, Zagoruyko et al. [78] conducted an extensive study on ResNet block structures and proposed a new design that reduces network depth while increasing its width, resulting in Wide Residual Networks (WRNs). WRNs outperform their thin and very deep counterparts, and even a simple 16-layer deep residual network can achieve state-of-the-art results on ImageNet, surpassing thousand-layer deep networks in both accuracy and efficiency.

Residual neural networks (ResNets) have proven to be effective in various medical imaging tasks, including classification, segmentation, and detection, leading to significant progress in the field of intelligent medicine. In medical imaging classification and detection, ResNets can capture intricate and abstract features, resulting in superior accuracy compared to traditional machine learning approaches.

In the field of medical image segmentation, Li et al. [79] proposed a 3D convolutional network architecture for brain segmentation that utilizes dilated convolutions and residual connections to capture a large volume of context. Their network is more conceptually simple and compact compared to state-of-the-art segmentation networks. In a similar vein, Alom et al. [38] introduced RU-Net and R2U-Net, which are recurrent and recursive residual convolutional neural networks based on the U-Net model. These models combine the capabilities of U-Net, Residual Networks, and RCNN, resulting in improved feature representation through feature accumulation with recurrent residual convolutional layers. They also have the same number of parameters as U-Net, making them a more efficient architecture. These models have outperformed other models, including U-Net and Residual U-Net, on benchmark datasets for medical image segmentation.

To summarize, residual networks have been found to enhance network performance, increase depth, and improve the network’s resilience to problems like model degradation. Moreover, the skip connections used in residual networks do not add to the computational complexity during training. Building upon the work of previous researchers, this thesis proposes an improved method for utilizing residual blocks in U-Net. By incorporating skip connections in the residual block, this method addresses the information loss resulting from deepening U-Net.
CHAPTER 2. RELATED WORK

2.3 Attention mechanism

The concept of attention has become a crucial element in deep learning, inspired by the ability of the human visual system to naturally identify important regions in complex scenes. This mechanism emphasizes significant local regions while filtering out irrelevant information, replicating the biological process, and aiding in the modeling of long-term dependencies. Additionally, attention mechanisms can be used as a resource allocation scheme and a primary means of addressing the problem of information overload. Due to the advances in deep neural networks, attention mechanisms have become widely used. For example, Mnih et al. [50] developed an attention model that adaptively selects a sequence of regions or locations for processing. Meanwhile, Chen et al. [80] trained several attention masks to fuse feature maps or predictions from different branches. Vaswani et al. [81] learned a self-attention model for machine translation. Xu et al. [82] and Yao et al. [83] have utilized attention models for image and video captioning tasks, respectively.

2.3.1 The combination of attention mechanism and U-Net

In recent years, attention mechanisms have been successfully integrated into medical image segmentation, with notable achievements [45, 84]. The U-Net architecture involves progressively recovering downsampled images through pooling and strided convolution. Incorporating attention mechanisms into U-Net can facilitate information flow from deep to shallow layers, thereby improving the effectiveness of upsampling. Moreover, U-Net can be tailored to focus on specific regions of interest in an image, allowing the model to emphasize crucial features and attenuate irrelevant ones. Thus, by enhancing the model’s ability to capture local responses, segmentation accuracy for medical images can be improved.

Medical imaging has seen promising results with the use of Attention U-Net in various applications such as organ, tumor, and lesion segmentation. Multiple studies have shown its effectiveness, with Attention U-Net achieving state-of-the-art results in brain tumor segmentation in the BraTS challenge and showing promise in the segmentation of prostate, liver, and lung tumors.

In one such study, Zhang et al. [85] aimed to investigate the impact of locally important responses on brain tumor segmentation by introducing the attention mechanism into the existing U-Net architecture. They proposed the Attention Residual U-Net (AResU-Net), an end-to-end two-dimensional network that simultaneously incorporates attention mechanisms and residual blocks into U-Net to further improve segmentation accuracy. AResU-Net enhances the local response of downsampling residual features and facilitates feature recovery in the subsequent upsampling process by adding a series of attention units in the corresponding downsampling and upsampling process, resulting in adaptive feature adjustments. Similarly, Li et al. [86] developed ANU-Net, an attention-based nested segmentation network, that includes an attention mechanism between nested convolutional blocks to integrate features extracted at different levels, resulting in improved segmentation across various organs in common medical imaging modalities. Awasthi et al. [87] introduced a multi-threshold model based on Attention U-Net that identifies distinct regions of glioma in MRI scans with low computational complexity, memory requirements, and training time. Aboelenein et al. [88] proposed Multiple Initial Residual Attention U-Net (MIRAU-Net), a novel codec architecture that integrates residual and initial modules into U-Net to improve glioma segmentation performance. The codecs in this architecture are connected via initial residual paths to reduce the distance between their feature maps. The MIRAU-Net architecture delivers superior performance, particularly for small-scale brain tumors.
Attention U-Net has shown promise in segmenting various structures in medical imaging beyond tumors, making it a versatile approach with potential applications in the diagnosis and treatment of various medical conditions. Oktay et al. [40] introduced a novel attention gate (AG) model for medical imaging, as shown in Figure 2.4, which can be easily integrated into the U-Net architecture with minimal computational cost while improving model sensitivity and prediction accuracy. Chen et al. [89] introduced an improved U-Net model for accurate quantification of COVID-19 infection measured by computed tomography (CT) images, utilizing a residual network to enhance feature extraction and an effective attention mechanism in the decoding process.

Attention U-Net has demonstrated significant potential in medical image segmentation, with multiple studies showcasing its effectiveness. This fusion of U-Net architecture and attention mechanism not only retains the strengths of the original U-Net structure but also improves its performance by allowing it to focus on specific regions of interest. As a result, Attention U-Net has emerged as a powerful tool for computer vision applications in medicine, offering promising opportunities for accurate and efficient diagnosis, treatment, and monitoring of various medical conditions.
Chapter 3

Methodology

Convolutional neural networks (CNNs) are known for their ability to extract local features through the use of convolutional filters, but they may struggle with capturing long-term relationships and dependencies between different objects in an image. This limitation can lead to the misidentification of similar features and hinder the model’s performance. To address this issue, it’s important to incorporate long-term semantic information and context dependencies into the weight values of the model. Failing to do so can result in overfitting and model degradation. The U-Net architecture is an artificial neural network that is specifically designed for image segmentation tasks. Its symmetric encoder-decoder structure consists of multiple convolutional layers that work together to improve the accuracy of the model.

To address the limitations of UNet, we propose a new architecture called ArUnet. It builds upon the U-shaped architecture of U-Net but replaces the original skip connection with an attention connection to capture contextual information through encoders at the depth of each layer and generate masks by the decoders. By integrating U-Net and attention mechanism, ArUnet can better model long patch dependencies and encode the relative positional relations of different features. In addition, the encoders learn features via spatial dimensions on the contracting path, which are then fed into the attention connection to support the decoders on the expansive path. This allows for the fusion of features extracted separately from CNN filters and attention mechanism, resulting in improved depth of feature capturing while maintaining the original stability of UNet. The use of residual blocks also greatly reduces the possibility of model degradation. Overall, ArUnet provides a promising solution for improving the performance of image segmentation tasks.

3.1 Overall Structure of ArUnet

In recent years, there has been a growing interest in combining self-attention mechanisms with the U-Net architecture to improve performance on image segmentation tasks. Typically, these models use the U-Net architecture as the backbone and incorporate self-attention mechanisms at various stages of the network. By doing so, these models can capture long-range dependencies and better contextual information, which can improve the accuracy of image segmentation.

To incorporate the self-attention mechanism into the U-Net architecture, our thesis introduces a novel modification, which involves using self-attention to guide the skip connections in the U-Net architecture. Specifically, a self-attention block is used to convert the feature representation in the encoder into attention weights. These attention weights are then used to compute skip connections
CHAPTER 3. METHODOLOGY

that are fused into the decoder for feature learning. This approach allows the model to selectively incorporate low-level and high-level features based on their relevance to the segmentation task. By using self-attention to guide the skip connections, the model can better capture long-range dependencies and contextual information, which can improve the accuracy of image segmentation.

3.1.1 U-shape Architecture

The U-Net architecture is a widely-used neural network design for image segmentation tasks, recognized for its distinctive encoder-decoder structure. The encoder module is composed of several convolutional layers that progressively decrease the spatial resolution of the input image, while the decoder module consists of upsampling layers that gradually increase the spatial resolution back to the original size. Skip connections are utilized to establish connections between corresponding layers in the encoder and decoder, allowing the model to retain both low-level and high-level features. This characteristic makes the U-Net architecture especially suitable for tasks in which precise localization of object boundaries is required, as it allows the network to integrate local and global information in a more effective manner.

The framework proposed in our work, depicted in Figure 3.1, combines a contraction path for context capture and an expansion path for precise positioning. To enhance the effectiveness of the learned features, the network adopts an attention mechanism and residual function. Specifically, an attention mechanism is embedded in each skip-connection layer of the U-Net architecture, allowing the model to selectively emphasize informative features. Furthermore, the entire network uses a residual function to preserve resolution and rich high-level feature information, without adding more parameters. This approach results in a compact yet powerful network that achieves improved segmentation performance.

A convolutional block, which includes a $3 \times 3$ convolutional layer, an instance normalization layer, and a ReLU activation layer, is utilized in the architecture. Downsampling is performed using stride-2 convolutional blocks instead of max pooling, and the number of feature channels is doubled.

Figure 3.1: Architecture of ArUnet.
in each downsampling layer. Upsampling involves an upsampling layer followed by a convolutional block, and the number of feature channels is halved in each upsampling layer. To capture more accurate contextual information, the feature maps from the upsampling layer are recombined with the feature maps learned from the remaining attention at the same level through a concatenation operation. The final output is generated using a softmax activation function.

3.2 Residual Modules for Down-sampling and Up-sampling

The improved U-Net model used in this thesis gradually deepens as the functionality increases. Therefore, ensuring the model's resistance to degradation is our top priority in conducting experiments. To address this issue, we chose to use residual structures to mitigate the randomness of gradient descent, obstacles to feature learning, and damage to image spatial structure. The mathematical meaning of residual structures will be explained in detail in the following text.

3.2.1 The function of residual blocks

ResNet's use of residual functions and skip connections allows the network to effectively identify and isolate significant features from the input, while also facilitating more efficient gradient propagation across the network. This design ensures that the model's performance remains consistent even as the architecture becomes deeper, effectively mitigating the vanishing gradient problem common in deep neural networks and promoting more rapid and stable training.

In each Residual Network unit, the input and output are denoted by $x$ and $y$, respectively, and the block comprises a residual function $F$ and parameters $W_i$. These residual blocks are then stacked in sequence to construct the Residual Network.

$$y = F(x, \{W_i\}) + x$$ (3.1)

In this context, the vectors $x$ and $y$, are the input and output of the respective layers under consideration. The function $F(x, W_i)$ is the residual mapping.

The dimensions of $x$ and $F$ in Equation 3.1 are generally expected to be identical. However, when this is not the case - such as when convolutional layers alter the input and output channel dimensions - a linear projection $W_s$ can be used via a shortcut connection to align the dimensions, as shown in Equation 3.2.

$$y = F(x, \{W_i\}) + W_s x$$ (3.2)

The vectors $x$ and $y$, are the input and output of the respective layers under consideration. The function $F(x, W_i)$ is the residual mapping. Linear projection $W_s$ is used to align dimensions.

The flexibility of the residual function $F$ is evident, allowing it to be utilized in a wide range of network architectures, from those with just two layers to much deeper models. Furthermore, the $F(x, W_i)$ mapping has the flexibility to represent multiple fully connected layers or convolutional layers. By introducing a deep residual learning framework to address the problem of degradation, the framework focuses on fitting residual maps rather than the underlying maps of all stacked layers. To achieve this, the network employs a shortcut connection in Equation 3.1 that introduces neither additional parameters nor computational complexity, making it practical for use in real-world applications. This connection ensures that the modified model retains the same number of parameters, depth, width, and computational cost as the original model.
The feedforward neural network with 'shortcut connections' shown in Figure 3.2 can achieve the residual function of $F(x) + x$. These shortcut connections can skip one or more layers and often perform a direct mapping by summing their outputs with the stacked layers of the original network. Despite the addition of shortcut connections, these networks can still be trained end-to-end using SGD with backpropagation and can achieve improved results. In this thesis, due to the inclusion of numerous convolutional layers and upsampling layers in the encoder and decoder structures, there is a higher likelihood of encountering the issue of model degradation during the experimental process. Additionally, an excessive number of convolutional layers can lead to unnecessary feature loss, which is detrimental to model training. Therefore, the characteristics of skip connections in residual networks are utilized to mitigate gradient descent and prevent model degradation. Simultaneously, at the end of the skip connections, shallow-level features are integrated into deep-level features to minimize unnecessary information loss and enhance the model’s resistance to depth and robustness.

### 3.2.2 Residual connection

In our architecture, we utilize residual blocks instead of MaxPooling to perform downsampling and maintain learned features in the process. Each downsampled residual block consists of three convolutional layers, each followed by a batch normalization layer and a ReLU activation function, as shown in Figure 3.3. The principle of the residual block is to introduce a residual connection, which greatly improves the training speed and accuracy without adding any additional parameters. The residual operation is expressed as $y = F(x, W_i) + x$, where $x$ and $y$ are the input and output vectors of the relevant layer, and $F(x, W_i)$ is the mapping function of the residual path. However, since the downsampling process geometrically reduces the input, the skip connection needs to be modified accordingly, that is to use a convolutional layer to ensure that the output sizes are equal.

The purpose of the upsampling module for feature recovery is to restore the size of the image layer by layer, restore the shallow features lost in the downsampling process, and use the upsampling
module to extract fusion features. In order to enhance the decoding performance, we introduce a similar residual structure after upsampling and feature splicing. The basic structure of the upsampling residual block includes two convolution units and a skip connection, as shown in Figure 3.3 which can perform element-wise addition operations on the input feature map and the convolution result to complete the residual function.

The residual blocks are powerful tools that can improve the decoding performance of our architecture. By connecting input feature maps as a shortcut connection, the feature information of the input can be maintained. Then, the element-wise addition operation combines shortcut connections with residual paths, which can better preserve feature information, especially in deep models. The decoder module outputs a mask of the same size as the original input. By using the combination of several Res blocks, we can recover richer feature information and achieve better image restoration results through deconvolution.

3.3 Feature Fusion via Self-Attention

This thesis proposes a novel way to fuse attention-learning mechanisms between convolutional neural networks. The essence of the attention mechanism is to learn the weights of important features by establishing a shadow matrix, and then increasing the weights of these important features using the principle of vector multiplication for weight allocation. Ultimately, the convolutional network is employed to fuse the important features, accomplishing the purpose of learning. Fusing attention mechanisms into convolutional channels will further enhance the feature extraction ability of convolutional kernels, and at the same time, strengthen the convergence effect of the entire model’s nonlinear fitting. Due to the unique calculation of attention mechanisms, the generated model structure coincides with the decoder structure in the U-Net network, which uses deep features to enhance
the learning effect of shallow features and preserves the spatial structure of the image on the final output using shallow features. The following section will focus on the calculation of attention mechanisms and the generation of model structures.

Figure 3.4: Self-attention [81].

3.3.1 Matrix Q, K, V of Self-attention

As we encounter scenes in our daily lives, our visual attention system is naturally drawn to specific regions, allowing us to process relevant information quickly. This cognitive process can be described as:

\[ \text{Attention} = f(g(x), x) \quad (3.3) \]

In this context, the function \( g(x) \) generates attention by identifying discriminative regions that warrant further processing. The resulting attention, denoted by \( f(g(x), x) \), is used to selectively process the input \( x \), akin to focusing on key regions and extracting pertinent information.

The definition outlined above suggests that most current attention mechanisms can be formulated using the same basic equation. It’s worth noting that a function can be defined as a mapping between a query and a set of key-value pairs, where each of these elements (\( Q, K, V \), and the output itself) is represented as a vector. To compute the output, the values are combined using a weighted sum, where the weight assigned to each value is determined by a compatibility function between the corresponding key and the query.

As an illustration, we will use the example of self-attention. For self-attention, the functions \( g(x) \) and \( f(g(x), x) \) can be defined as follows:

\[ Q, K, V = \text{Linear}(x) \quad (3.4) \]

\[ g(x) = \text{Softmax}(QK) \quad (3.5) \]
Self-attention techniques are employed to compute contextual encodings for positions in a sentence by taking a weighted sum of all positional embeddings, as shown in Figure 3.4. Based on this figure, the model structure can be generated. Three weight matrices are obtained through matrix transformations, and then they are treated as inputs to the convolutional network. The final result is obtained using matrix calculations, which enhances the distribution of the image’s feature structure in this result.

### 3.3.2 Computing The Attention in Convolutional Neural Networks

The ArUnet architecture employs residual blocks in both the network’s encoding and decoding stages, enabling the network to maintain spatial information and minimize the loss of feature maps during the downsampling and upsampling operations. Meanwhile, attention mechanisms, particularly self-attention, have demonstrated their ability to capture long-range dependencies in feature maps, resulting in the better overall performance of networks.

In this thesis, we also replace skip connections in the neural network architecture with a self-attention mechanism that selectively focuses on relevant feature maps while suppressing irrelevant ones. The self-attention mechanism computes an attention map based on the input feature maps, which weighs the importance of each feature map. Higher weights indicate greater importance for the current task, enabling the model to selectively focus on feature maps that are beneficial for the task. Simultaneously, it suppresses irrelevant feature maps, those that are unrelated or less important.
to the task, by reducing their weights to minimize or disregard their impact on the model. To achieve this, a learned weight matrix is applied to the input feature map to produce a set of attention coefficients, which are then normalized using a Softmax function to ensure they sum to one. The coefficients are then used to weight the feature maps before passing them to the next layer. This approach enables the network to effectively capture long-range dependencies in feature maps and improve its overall performance.

Incorporating the attention mechanism in place of skip connections in U-Net can improve the performance of the network on image segmentation tasks. By selectively focusing on relevant feature maps and suppressing irrelevant ones, attention mechanisms can help to enhance the accuracy and robustness of segmentation models. This results in a more efficient network that is better able to extract meaningful features from input data, leading to improved segmentation performance.

In the architecture we proposed, as illustrated in Figure 3.1, the encoder and decoder paths are constructed using residual blocks with residual connections. Each residual block includes a convolutional layer, an instance normalization layer, and a ReLU activation layer. Subsampling is performed using stride-2 convolution blocks instead of max pooling, and the number of feature channels is doubled in each subsampling layer. Upsampling involves an upsampling layer and a residual convolution block, with the number of feature channels halved in each upsampling layer. To further enhance the performance of the network, several residual blocks are connected to the subsampling and upsampling modules respectively to extract and fuse features. Attention blocks are used in the encoder and decoder paths before the combination of low-level and high-level features to balance the importance of different features. Finally, the output feature map is normalized and fed into the output layer. This approach maximizes the preservation of spatial information while achieving effective feature extraction and fusion, thereby improving segmentation accuracy.
Chapter 4

Experiment and Results

The main objective of this study is to enhance the performance of medical image segmentation using a modified UNet architecture with an attention mechanism. To achieve this goal, we have made several efforts in the following aspects: Firstly, we have optimized the procedure of data processing and improved the noise reduction effect. Secondly, we have enhanced the stability of the model to improve the experimental outcomes. Lastly, we have combined various machine learning evaluation methods to optimize the model’s output.

4.1 Dataset and Data Processing

The segmentation of brain tumors in MRI images is a challenging task in medical image analysis due to the complexity of brain structure, biological organization, and the impact of imaging quality. While deep learning models such as the U-Net network are known to be robust to noise, data processing is still a crucial step to improve their performance. In this work, we focused on a multimodal MRI brain scan dataset, specifically the BraTS 2021 dataset (as shown in Figure 4.1), and utilized the U-Net network for brain tumor segmentation.

The BraTS 2021 dataset provides a very large and comprehensive annotated database of brain MRI scans with detailed labeled fragments. It also contains 1251 scans with ground-truth annotations of tumor regions and includes 4 MRI scan modes: native (T1), post-contrast t1-weighted (T1Gd), t2-weighted (T2), and FLAIR. For each case, the ground truth segmentation includes four classes: enhancing tumor (ET) (label 4), peritumoral edema (ED) (label 2), necrotic tumor core (NCR) (label 1) and background (label 0). The tumor core (TC) is defined as the sum of ET and NCR, and the whole tumor (WT) consists of the sum of ED, ET and NCR. The labels provided are annotated by one to four annotation officers and reviewed and approved by neuroradiologists. In our experiments, we chose t2-weighted (T2) with necrotic tumor core (NCR) (label 1) for model training and verification. T2 shows better tissue lesions, label 1 is to identify the tumor core.

To optimize the network’s performance, we performed preprocessing on raw MRI brain tumor images. This included normalization, which helped to improve the quality of the images and remove any irrelevant information that could affect the segmentation results. By taking these steps, we were able to improve the accuracy of the U-Net network’s predictions and obtain more precise segmentation results.
Figure 4.2 illustrates the various steps we took for data processing. Firstly, we selected the slice with the highest feature density from the original 3D brain image data. The method is based on the measurement of cross-sectional area, aiming to include as many tumor features as possible within the 2D slices. We then segmented each 3D image into 2D image slices and removed unnecessary pixels such as image edges from each slice. This resulted in an image size of 128 × 128, which was then normalized using z-score normalization to achieve zero-mean normalization. Specifically, we calculated the z-score normalization as $z_0 = (z - m)/d$, where $z$ and $z_0$ are the input and normalized output images, respectively, and $m$ and $d$ are the mean and standard deviation of the input image, respectively. By following these steps, we were able to preprocess the data to remove any irrelevant information and improve the quality of the images.
4.2 Training Procedure and Evaluation Metrics

To comprehensively evaluate the performance of our proposed image segmentation framework, we adopted four commonly used metrics in this field: the Dice coefficient, sensitivity, specificity, and Hausdorff distance. These metrics are particularly relevant for assessing the accuracy of the U-Net neural network. The Dice coefficient measures the degree of overlap between the predicted and ground truth segmentations, calculated as twice the number of common voxels divided by the sum of voxels in both masks. Sensitivity and specificity measure the proportion of true positives (TP) and true negatives (TN) correctly identified, respectively. The Hausdorff distance quantifies the maximum distance between the predicted and ground truth masks, determined by computing the distance between all points on each mask and finding the maximum value. We evaluate these metrics separately for each of the three subregions of interest. The corresponding formulas for calculating these metrics are as follows:

\[
\text{Dice coefficient} = \frac{2 \times TP}{2 \times TP + FP + FN} \quad (4.1)
\]
\[
\text{Sensitivity} = \frac{TP}{TP + FN} \quad (4.2)
\]
\[
\text{Specificity} = \frac{TN}{TN + FP} \quad (4.3)
\]
\[
\text{Hausdorff distance} = \max(d(p, t), d(t, p)) \quad (4.4)
\]

where TP, FP, TN, and FN denote true positives, false positives, true negatives, and false negatives, respectively. The letters p and t represent the voxels in the predicted segmentation P and ground truth segmentation T, respectively. We calculate the distance between p and t as \(d(p, t)\) and the distance between t and p as \(d(t, p)\) using the Euclidean distance. These metrics are able to accurately evaluate the performance of our framework in segmenting three sub-regions of interest in medical images. By utilizing these measures, we can effectively assess the accuracy of our proposed framework and compare it to other state-of-the-art methods in the field of medical image segmentation.

4.3 Implementation

We implemented our proposed deep learning models and frameworks using the PyTorch library. Furthermore, to avoid errors caused by changes in the experimental environment, we reimplemented several published models using PyTorch and applied a control variable approach for more comprehensive and controlled experiments. To improve the efficiency of our training process, we uploaded the dataset to Google’s hard disk for processing and conducted all experiments using the Google Colab environment.

Our dataset consists of a total of 1,251 brain tumor MRI images, of which 850 images are used as the training dataset and the remaining 401 images are used as the validation dataset. In training, we tested a hybrid loss function comprising binary cross-entropy loss, Dice loss, and a combination of both as our objective function. We optimized the model using SGD and Adam optimizers. To comprehensively evaluate the model’s performance, we employed multiple indicators such as Dice coefficient and Hausdorff distance as auxiliary evaluation criteria. After conducting extensive experiments, we found that using the combination of Dice loss and Adam optimizer yielded the best results.
Figure 4.3: The performance of the trained ArUnet on the verification data set.
Figure 4.4 shows a scatter plot of the accuracy and loss function achieved on the validation dataset, and the fitted curve shows that the model converges well. In addition, Figure 4.4 displays the scores obtained from each of the four evaluation methods on the test dataset. DiceScore, calculated using Dice loss with the highest score of 95.54%. Sensitivity and Specificity, have both achieved an accuracy rate of over 90%, providing an objective representation of the model’s fit. Finally, the Hausdorff distance measures the difference between two images. As the number of epochs increases, the Hausdorff distance gradually decreases from the initial 87.0775 to 13.8128, accurately depicting the model’s performance improvement during training.

4.4 Results and Analysis

To evaluate the effectiveness of our proposed architecture and compare its performance with other popular architectures, we designed two sets of experiments. In the first set, we compared our architecture, ArUnet, with a standard UNet, a residual UNet, and a attention U-Net without residuals. However, due to the deep U-Net architecture and the size of the BraTS2021 dataset, attention-based networks without hybrid residual blocks exhibited model degradation, which is lead to a sudden decrease of the model performance after a particular training period. In the second set of experiments, we compared our model with various other attention-based U-Net architectures.
Based on our experimental results, our proposed architecture outperforms standard UNet, residual UNet, and other attention-based UNets in terms of segmentation accuracy. Firstly, the underlying U-Net model achieved a result of 92.8% on the BraTs dataset. Subsequently, the fusion of the self-attention mechanism increased the result to over 93%. Further experimental results demonstrated that the addition of the residual network significantly improved the retention of spatial features,
which is crucial in improving the performance of image segmentation. This improvement was observed with an increase in accuracy of nearly three percent on the original basis. According to these results, we optimized our proposed ArUnet architecture, which achieved 95.54% accuracy on the validation dataset. Therefore, we discovered that incorporating the attention mechanism into the U-Net architecture improves its segmentation performance.

Experiment 1 provided further insight into the influence of residual networks and attention mechanisms on model training. To analyze this, we plotted the loss curves for the four models used in the first set of experiments. As illustrated in Figure 4.5, the network with residual connections demonstrated more stable convergence, thereby enhancing the model’s resistance to degradation. Additionally, the incorporation of the attention mechanism significantly improved the model’s performance compared to the standard network. Figure 4.6 displays the convergence speed and best scores of the different models under the agreed training parameters, which more explicitly emphasizes the attention mechanism’s ability to focus on key features.

For the second experiment, we compared our proposed ArUnet model with several of the well known advanced models. We tried multiple sets of hyperparameter combinations, compared learning rates, batch numbers, and selected the best parameters for each model. For example, the learning rate is compared as follows, with several typical values such as 0.1, 0.01, 0.001, 0.0001, etc., and 0.001 is confirmed to be the optimal learning rate. The batch number is selected from 1 to 64, with a multiple of 2 as the step size, and the result is 1 as the best. Both AttUnet and SwinAttUnet used the same idea as the structure proposed in this thesis, that is, to integrate the self-attention mechanism to improve the ability of feature extraction. UTNet utilized the Transform structure to enhance the U-Net decoder's ability, fully integrating its encoding and decoding structure. By comparing the segmentation performance on the tumor core, based on the same comparison benchmark, the accuracy rate of AttUnet and SwinAttUnet failed to reach 94%, but the accuracy of UTNet can reach 95.49%. However, despite being one of the most advanced models, UTNet’s accuracy was slightly lower than the structure proposed in this thesis. Figure 4.7 displays the convergence speed.
and the best score of different models under the respective training parameters that can achieve the best performance, which more clearly highlights the ability of the attention mechanism to focus on key features. Overall, observing the curve trend of the four models, that is, the slope of the fitting curve, the structure proposed in this thesis slightly outperformed UTNet on the verification dataset, achieving 95.54% accuracy. UTNet’s idea of integrating the Transform decoder with the U-Net decoder can further improve image segmentation performance in the future, based on the structure proposed in this thesis.

<table>
<thead>
<tr>
<th>Network</th>
<th>Accuracy</th>
<th>Loss</th>
<th>Epoch</th>
</tr>
</thead>
<tbody>
<tr>
<td>U-Net</td>
<td>0.9282</td>
<td>0.07182</td>
<td>61</td>
</tr>
<tr>
<td>AttUnet</td>
<td>0.9305</td>
<td>0.06957</td>
<td>40</td>
</tr>
<tr>
<td>SwinUnet</td>
<td>0.938</td>
<td>0.062</td>
<td>69</td>
</tr>
<tr>
<td>ResUnet</td>
<td>0.9534</td>
<td>0.04667</td>
<td>29</td>
</tr>
<tr>
<td>UTNet</td>
<td>0.9549</td>
<td>0.04516</td>
<td>36</td>
</tr>
<tr>
<td>ArUnet</td>
<td><strong>0.9554</strong></td>
<td><strong>0.04459</strong></td>
<td>55</td>
</tr>
</tbody>
</table>

Table 4.1: Summary of Experimental Results.

To enable a more intuitive comparison of experimental results, we summarized the model data from all experimental control groups. Figure 4.8 shows the comparison of segmentation results of each model. Table 4.1 demonstrates that ArUnet exhibited more stable and better segmentation performance across multiple experiments. Furthermore, we recorded the number of epochs required for each model to achieve optimal performance. The results indicate that each model can reach an optimal value, typically between 30 and 70 epochs. In this experiment, the top-performing models were residual Unet, UTNet, and ArUnet. As the accuracy of these models improved during training, it became evident that the number of epochs required for convergence increased, suggesting that the model’s complexity affects its convergence speed. This underscores the importance of optimizing the network structure, minimizing training parameters, and improving training speed. Moreover, adding layers to the feature extraction process could potentially enhance the model’s image segmentation performance, which is a critical area for future research on the proposed model in this thesis.
Figure 4.8: The comparison of segmentation results of each model.
Chapter 5

Conclusion and future works

The analysis and processing of medical images have a significant impact on clinical applications and scientific research. New ideas for medical image analysis can be provided by deep learning, allowing for the detection of morphological or texture patterns solely from the data. The main goal of brain tumor medical image analysis has become the improvement of deep learning methods due to the significant progress made on different medical imaging systems. Therefore, based on existing research, an improved UNet-based model is proposed. Firstly, extensive research was conducted and the existing methods for image segmentation were analyzed and classified differently. Two methods based on the encoder-decoder structure, the U-Net structure and self-attention mechanism, were selected. From this research, two important rules that form the theoretical basis for the proposed structure in this thesis were summarized. The first rule states that the U-Net structure can preserve deep and shallow features and merge them, providing a natural advantage for image segmentation tasks. The second rule indicates that the self-attention mechanism can enhance the ability to extract important features.

A new architecture called ArUnet is proposed in this thesis to achieve accurate segmentation of brain tumors. The self-attention mechanism is utilized to improve the skip connection in the U-Net structure, enhancing the performance of feature extraction in this channel and completing the fusion of deep and shallow features. Additionally, to address the issue of model degradation as the complexity increases, residual connections from the ResNet structure are employed to ensure the continuous increase in model depth. In the final completed model experiment, convolutional networks are organically integrated with attention mechanisms while retaining the advantages of both, resulting in performance that exceeds the current state-of-the-art network on the BraTS2021 dataset, achieving up to 95.54% accuracy.

However, the model’s expressive power is constrained by the limited computing resources available in Google Colab, which also limits its depth. Therefore, there are still many areas for optimization in the ArUnet proposed in this thesis. These areas include further optimization of the network structure to reduce the number of model parameters and increase training speed and positively impact the model’s stability. The robustness of the model to pathological images of different types of tumors or even diseases has not yet been validated, making it a focus of future work and a new research direction for further deepening the application of deep learning models in medical diagnosis. Further research will also promote the automation of medical diagnosis and make necessary contributions to human health.
Bibliography


