MRI Glioma detection and classification using vision transformers

Supervised by:

Dr. Mohamed ElKholy

Presented by:

Ahmed Wael Ahmed

Mostafa Marwan Mostafa

Rashed Mamdouh Abdelmoneam

Ahmed Youssef Ismael

Ahmed Mohamed Ayman

Hisham Gaber Fathy

Abdelrahman Mohamed Elsayed

Table Of Contents:

1 Introduction:

1.1 Medical Imaging

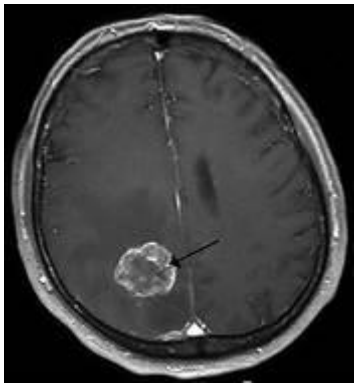
Medical imaging is a technique used to visualize the internal structures of the body for diagnostic and therapeutic purposes. Radiological imaging of the intact human body requires some form of energy to enter the body, interact, and then deliver a signal to a detector, typically located outside of the body. There are many types of medical imaging scans including but not limited to CT and MRI. Those are the most common medical images used in the field of AI.

1.2 MRI

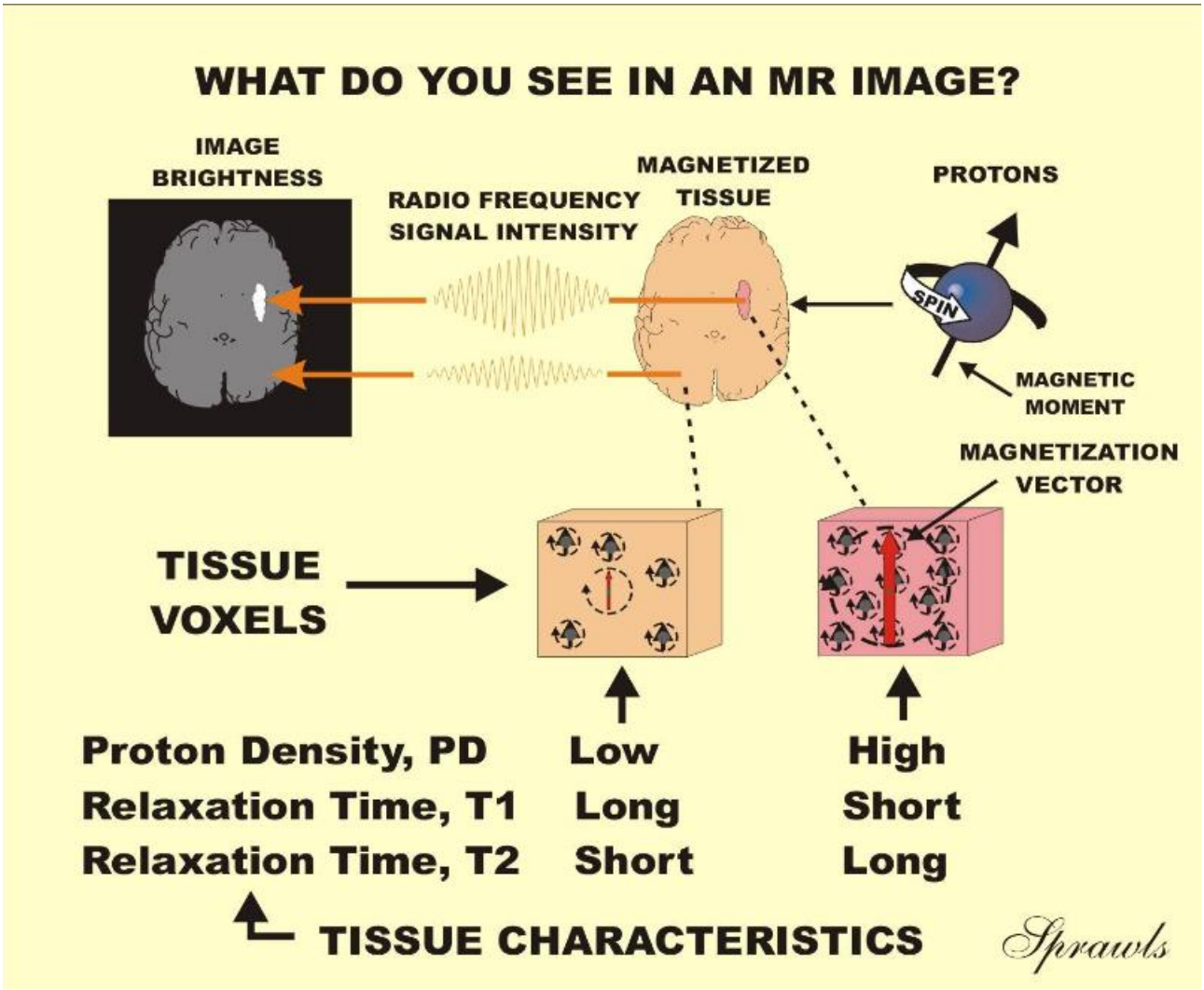
Magnetic Resonance Imaging is a type of medical scan that uses magnetic fields and radio waves to create detailed images of the body's internal structures, especially soft tissues. Unlike X-rays and CT scans, MRI does not use ionizing radiation, which makes it a safer option for patients. (They are explained in detail in the MRI attachment.)

1.3 MRI in brain tumors

MRI is an important tool used in the diagnosis and treatment of brain tumors. It helps doctors determine the size, location, and type of tumor, as well as monitor the effectiveness of treatment . It is also used to guide minimally invasive procedures such as MRI-guided laser ablation, which is a neurosurgical option for the treatment of brain tumors.



- How to detect Brain Tumor (Medical approach)



1.4 AI in medical imaging

Artificial Intelligence has revolutionized the field of medical imaging by providing a more accurate and efficient way of diagnosing and treating patients. AI algorithms can analyze medical images such as X-rays, CT scans, and MRI scans to efficiently detect abnormalities in the human body.

- How to detect Brain Tumor (AI Approach)

1. **Data Acquisition**: Collect a dataset of MRI images with labeled tumor and non-tumor regions. This dataset is crucial for training the AI model
2. **Preprocessing**: Clean and preprocess the MRI images. Steps may include resizing, normalization, and enhancement to improve model performance
3. **Segmentation**: Using Vision Transformer for image segmentation, separating the tumor region from the rest of the image
4. **Model Training**: Train a deep learning model, (Vision Transformer) using the preprocessed data. The model learns to identify patterns associated with tumors
5. **Validation and Testing**: Evaluate the model on a separate dataset to ensure its generalization. Testing involves predicting tumor presence and location in new, unseen images

1.5 Visual segmentation

Visual segmentation groups pixels of the given image into a set of regions. Each region represents a single entity in the image. Certain predetermined rules specify how the image is segmented. It is a fundamental problem in computer vision and involves numerous real-world applications, such as robotics, automated surveillance, image/video editing, social media, autonomous driving, etc. Over the past decade, deep learning-based methods have made remarkable strides in this area.

1.6 Semantic Segmentation

Semantic segmentation is the problem of assigning a class label to each pixel of an image. It is a fundamental topic in computer vision and is critical for various practical tasks. In Semantic Segmentation (SS), the classes may be foreground objects or background views like the sky, and each class only has one binary mask that indicates the pixels belonging to this class. Each SS mask does not overlap with other masks[19]. Global contextualization in SS refers to the process of capturing and integrating contextual information from image parts to improve the accuracy of segmentation [18][19][20].

1.7 CNNs

Convolutional Neural Networks are a type of Deep Learning neural network architecture commonly used in Computer Vision. They are primarily used to solve difficult image-driven pattern recognition tasks. CNNs are based on the principles of linear algebra, such as matrix multiplication, for detecting patterns in an image. They are made of a series of layers like convolutional and pooling layers. These layers consist of functions that modify the input image returning feature maps containing more important information relative to the model.

1.8 Transformers

Transformers are a type of neural network architecture that revolutionized the field of Natural Language Processing. They were introduced in a 2017 Google paper and have since been used in various applications in AI. The Transformer model is based on the attention mechanisms, which allows the model to have weighted attention (more focus on more relevant information).

1.9 Vision Transformers

We intend to use vision transformers in Computer Vision in the field of medical imaging. Transformer-based models perform similar to or better than other types of networks such as convolutional and recurrent neural networks[1]. Given its high performance, transformers are receiving more attention from the computer vision community. After the transformer was proposed, with the goal of global context modeling, several works design variants of self-attention operators to replace the CNN prediction heads (Transformer-Based Visual Segmentation).

Transformers' attention mechanisms proved to be better than CNNs in image segmentation and classification if large amounts of data are presented [4].

1.10 Global Contextual Modeling

Global Contextual Modeling is a mechanism that enables the network to capture long-range dependencies. It focuses on capturing information about the entire image or scene. This technique enhances the performance of semantic

segmentation by effectively utilizing the spatial and semantic relationships within an image[18][19][20].

1.11 Multi-Headed Attention

A core component of transformers, multi-headed attention allows the model to focus on different parts of the image simultaneously, capturing various features and relationships at multiple levels of abstraction[28].

2 Problem Statement (CNN limitations)

- **Limited Field of View:** CNNs have a limited receptive field, which means they might not capture global context or long-range dependencies in the data due to the naive structure in its building modules (local convolution kernel) [13][17].
- **Weak Scaling:** Although CNNs demonstrate optimal performance when applied to smaller datasets. Their scalability is ***relatively*** limited when there is an increase in the volume of the data for training [10].
- **Down sampling:** CNNs are frequently forced to implement down sampling or pooling layers to decrease the spatial size of feature maps to reduce memory consumption. This problem is highlighted because ***CNN is a sequential model*** (cannot train in parallel) [22].

3 CNNs Vs Transformers

Convolutional Neural Networks and Transformers are two popular deep learning architectures, each with its own strengths and applications.

- CNNs are designed to process data with a grid-like structure, such as images. They efficiently learn local patterns and spatial hierarchies in images through convolutional layers and pooling, reducing the dimensionality of input data while preserving critical information.
- Transformers, introduced in 2017, primarily used for natural language processing (NLP) tasks. The key innovation of Transformers is the use of self-attention mechanisms, which allow the model to weigh the importance of different parts of the input when making predictions. Transformers can effectively handle long-range dependencies and contextual information.

3.1 CNNs vs Transformers based on properties and architectures

References [10][22][23][20]

Points | Model	CNN	Transformer
Architecture	Sequential	Parallel
Inductive Bias	Have strong image-specific inductive biases, such as locality, two-dimensional neighborhood structure, and translation equivariance	Have much less inductive bias in small datasets but better inductive bias in larger datasets
Feature Extraction Nature	Local features due to the nature of its local kernels	Global feature due to the attention mechanism's nature
Performance with small data size	Better performance	Lower Performance (due to lack of bias)
Performance with low quality images	More Efficient and scalable (due to its fewer parameters)	Less Efficient (Excels in working with high dimensional images)

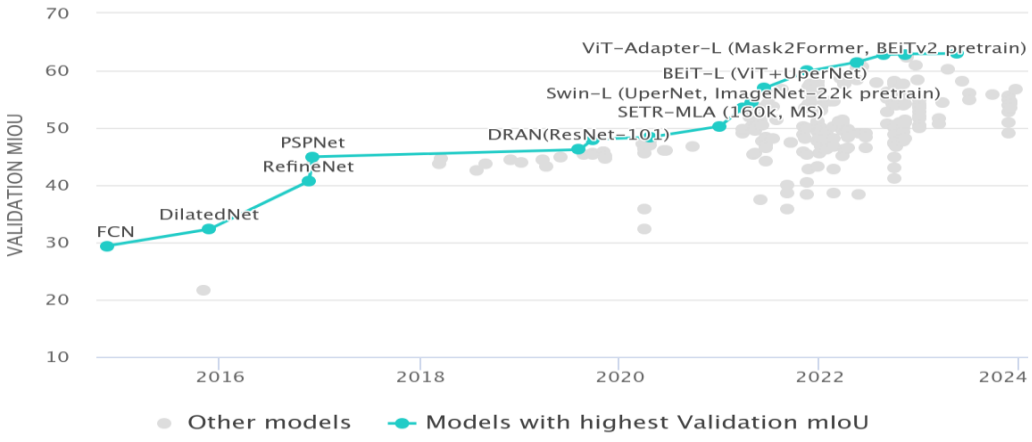
3.2 CNNs vs Transformers in Medical Images

In the field of medical imaging, both CNNs and Transformers have shown promise.

- CNNs have been the go-to approach to automated medical image diagnosis for a decade [23].
- Transformers, specifically Vision Transformers, have appeared as a competitive alternative to CNNs, yielding similar levels of performance while possessing several interesting properties that could prove beneficial for medical imaging tasks [24].

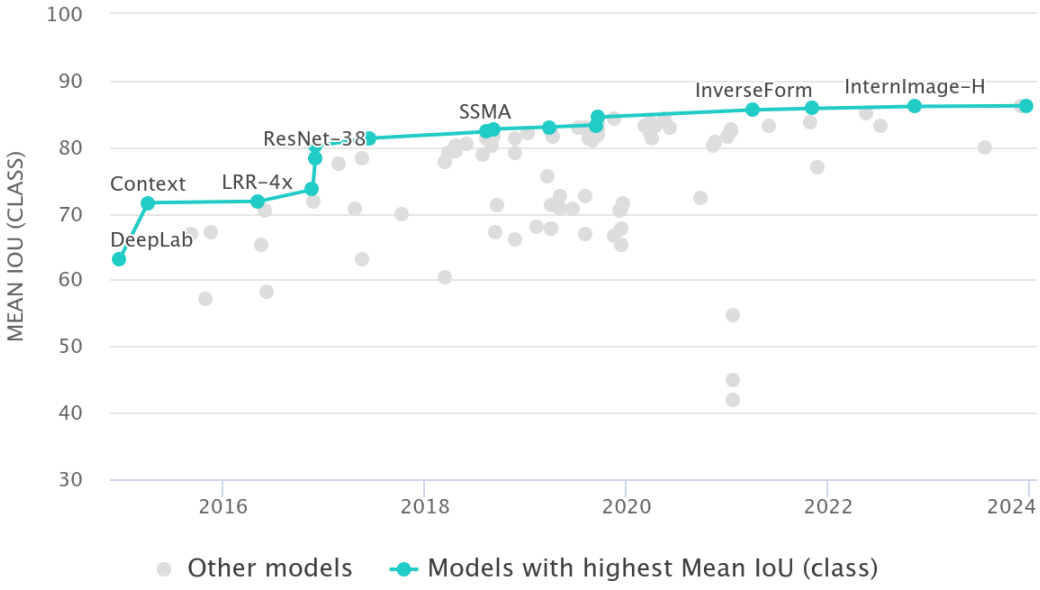
3.2 CNNs vs Transformers results in multiple competitions datasets

1- ADE20K



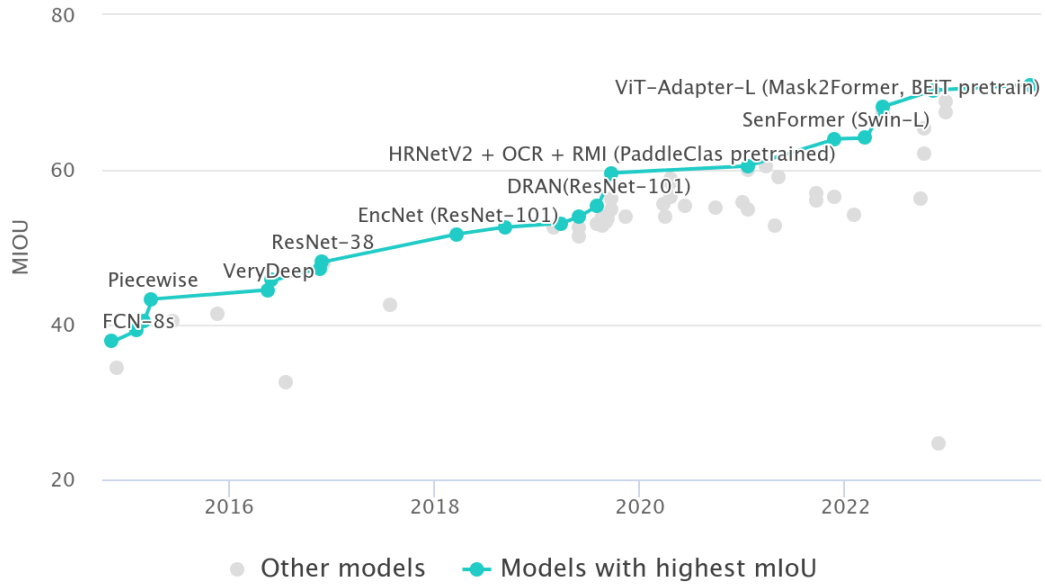
One of the best preforming models in this competition is Vit-adapter-L which is based on vision transformers. [25]

2- Cityscapes



The best model in this dataset is MetaPrompt-SD which is a variation of Query based segmentation transformers like DETR.[26]

3- PASCAL



The best model in this dataset is Plain-seg which is a variant of Vision transformers.[27]

In Conclusion: In the three datasets the IOU of models that use transformers are much higher than those that do not use transformers.

4 Aim

We aim to achieve state-of-the-art results with our Segmentation Transformer model which will be trained specifically on brain tumors datasets, to develop a model that is not only more accurate but also more robust to diverse types of Gliomas MRIs capable of handling the complexities inherent in medical imaging data. It will be combined with another model that is trained in classifying the Glioma degree/type. At the end all the results shall be delivered in the form of the segmented image with the tumor type classification.

5 Motivation

The field of medical imaging is a critical area where machine learning, specifically transformer models, can have a significant impact. Despite the success of transformers in natural language processing and computer vision, their application in the field of medical imaging is still in its early stages [14][10][22].

It has been found that in some cases RCNNs are better than vanilla CNNs in object recognition. RCNNs are better at capturing dependencies[15]. Then it has been found that in most cases transformers are better than RCNNs in object recognition. As transformers have self-attention (Capture long term dependencies)[16].

The primary motivation for this project is to address the limitations of CNNs and improve the efficiency of diagnosing and segmenting brain tumors to aid oncologists in their day-to-day job.

6 Possible Contribution

The resulting models will be trained on as realistic Glioma MRI images as possible to avoid bias in data. They will also be able to detect, segment and classify brain tumors into different types based on their characteristics.

7 Transformers in Segmentation Applications

While transformers are still relatively new in the field of computer vision compared to techniques like convolutional neural networks (CNNs), there are some examples where transformers have been used in image segmentation tasks:

1. **Facebook AI:**
    - Facebook AI Research (FAIR) has developed a framework called DETR (DEtection TRansformer) for object detection and segmentation. DETR utilizes transformers to directly predict object bounding boxes and class labels, eliminating the need for anchor boxes and post-processing steps typically used in traditional object detection.

2. Hugging Face:

- Hugging Face, a company specializing in natural language processing (NLP) models, has extended their expertise to computer vision with the release of Vision Transformer (ViT). ViT applies transformers to image classification tasks by dividing images into patches and processing them through transformer layers. While ViT is primarily designed for classification, it can potentially be adapted for image segmentation tasks.

3. OpenAI:

- OpenAI, the organization behind GPT (Generative Pre-trained Transformer) models, has explored the application of transformers in various domains, including computer vision. While their primary focus has been on language understanding, their research may pave the way for using transformers in image segmentation tasks.

4. Google Research:

- Google Research has investigated the use of transformers for image recognition and segmentation tasks. Their research on self-attention mechanisms, which are fundamental to transformers, may contribute to the development of transformer-based architectures for image segmentation.

5. Academic Research:

- Various academic researchers have explored the use of transformers in image segmentation tasks. For example, papers such as "Data-Efficient Image Recognition with Contrastive Predictive Coding" by Hénaff et al. and "End-to-End Object Detection with Transformers" by Zhu et al. propose transformer-based approaches for different computer vision tasks, including segmentation.

6. NVIDIA:

- One prominent example of a big company that utilizes transformers in image segmentation is NVIDIA. NVIDIA is renowned for its advancements in graphics processing units (GPUs) and artificial intelligence (AI) technologies, particularly in the field of computer vision.

- NVIDIA's research and development efforts have led to the creation of innovative models and frameworks for image segmentation tasks, leveraging transformer architectures. One notable example is the "Semantic Fovea" model, which was introduced by NVIDIA researchers in 2021. Semantic Fovea utilizes a hybrid architecture combining transformers with convolutional neural networks (CNNs) to achieve state-of-the-art performance in image segmentation tasks.

- The use of transformers in image segmentation allows for more efficient and effective processing of visual data, enabling tasks such as object detection, semantic segmentation, and instance segmentation. By harnessing the power of transformers, NVIDIA and other companies are pushing the boundaries of what is possible in computer vision applications, driving advancements in fields such as autonomous driving, medical imaging, and augmented reality.

NVIDIA Official Website

These are just a few examples where transformers have been applied or explored in the context of image segmentation. As research in this area progresses and transformer architectures continue to evolve, we can expect to see more applications and advancements in transformer-based image segmentation techniques.

Related Work

NestedFormer

TransAttUnet

Nested Former

Nested Modality-Aware Transformer for Brain Tumor Segmentation

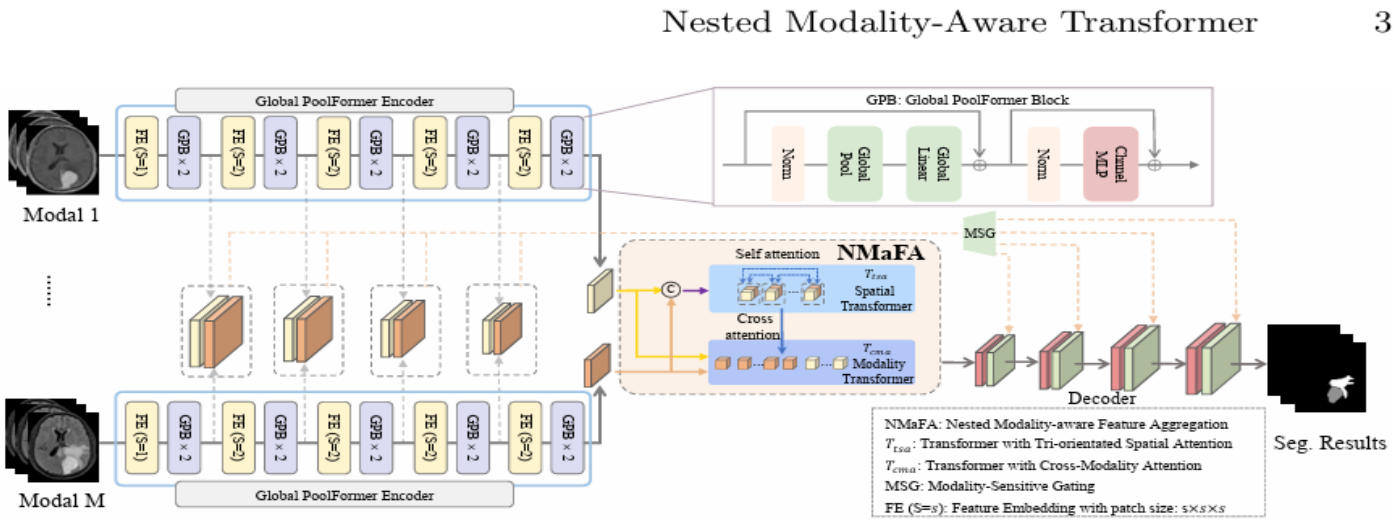
Intro

NestedFormer represents a novel approach to brain tumor segmentation within the domain of medical imaging. This method is tailored for multi-modal Magnetic Resonance Imaging (MRI), a standard diagnostic tool in clinical settings for the evaluation and diagnosis of brain tumors. The NestedFormer architecture features a multi-encoder single-decoder configuration, which has demonstrated superior performance relative to existing methodologies in extensive experiments using the BraTS2020 benchmark dataset. This dataset encompasses various MRI modalities, including T1, T2, and FLAIR sequences. Additionally, the performance of NestedFormer was validated on a private dataset dedicated to meningiomas segmentation, known as the MeniSeg dataset. The findings suggest that NestedFormer is a highly promising model for advancing the accuracy and reliability of brain tumor segmentation in medical imaging.

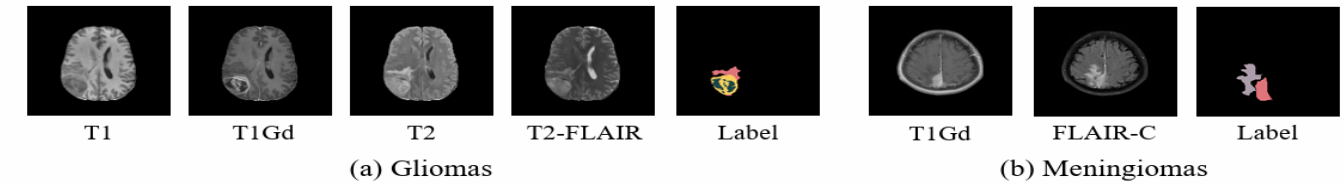
The key components of NestedFormer include:

1) Multiple encoders(Global PoolFormer Encoders)
2) Nested Modality-aware Feature Aggregation (**NMaFA**) module,
3) Modality-Sensitive Gating (**MSG**),

4) Single Decoder.



Types of Modals:



- From the images above, Nested former can aggregate multiple modals for the MRI image to make the segmentation required.

How it works:

1) Global Poolformer Encoder

Nested Former Overview:

- **Modality-Specific Encoders:** Utilizes multiple Global Poolformer Encoders, each corresponding to a different MRI modality (e.g., Encoder 1: T1 images, Encoder 2: T2 images, Encoder 3: FLAIR images).
- **Multi-scale Representation:** Captures features at various scales within each modality.

- **Global Pooling Mechanism:** Replaces traditional self-attention with global pooling to model global information, summarizing essential characteristics of the entire image for each modality, providing superior performance over self-attention methods.

2) Nested Modality-Aware Feature Aggregation (NMaFA) Module

Feature Combination:

- **Integration of Encoded Features:** Combines features extracted by the Global Poolformer encoders across different modalities.

Nested Transformers with Attention Mechanisms:

- **Tri-orientated Spatial Attention (Ttsa):**
  - **Spatial Attention:** Captures relationships between features within the same modality.
  - **Channel Attention:** Captures relationships between different feature maps within the same modality.
  - **Window-wise Attention:** Models relationships within local 3D windows.
  - **Position Encodings:** Utilizes learnable absolute position encodings for spatial and axial attention, and relative position encoding for window-wise attention, capturing long-range dependencies while maintaining computational efficiency.
- **Cross-Modality Attention (Tcma):**
  - **Cross-Modality Feature Interaction:** Concatenates features from different modalities along the spatial dimension.
  - **Cross-Attention Mechanism:** Explores relationships among modalities, enhancing modality dependency information within the features.

3) Modality-Sensitive Gating

Decoding Stage Focus:

- **Multi-modal Encoder Features:** Combines information from all modalities.
- **Modality Importance Maps:** Learns a separate importance map for each decoder layer.
- **Gating Mechanism:**
  - **Generation of Importance Maps:** Upsamples NMaFA module output and applies a fully connected layer followed by a sigmoid function, producing values between 0 and 1 indicating modality importance at specific locations.
  - **Feature Filtering:** Uses modality importance maps to filter encoder features during skip connections, ensuring the decoder focuses on the most informative data for each modality.

References for this part:

Xing, Z., Yu, L., Wan, L., Han, T., Zhu, L. (2022). NestedFormer: Nested Modality-Aware Transformer for Brain Tumor Segmentation. In: Wang, L., Dou,

Q., Fletcher, P.T., Speidel, S., Li, S. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. MICCAI 2022.

TransAttUnet

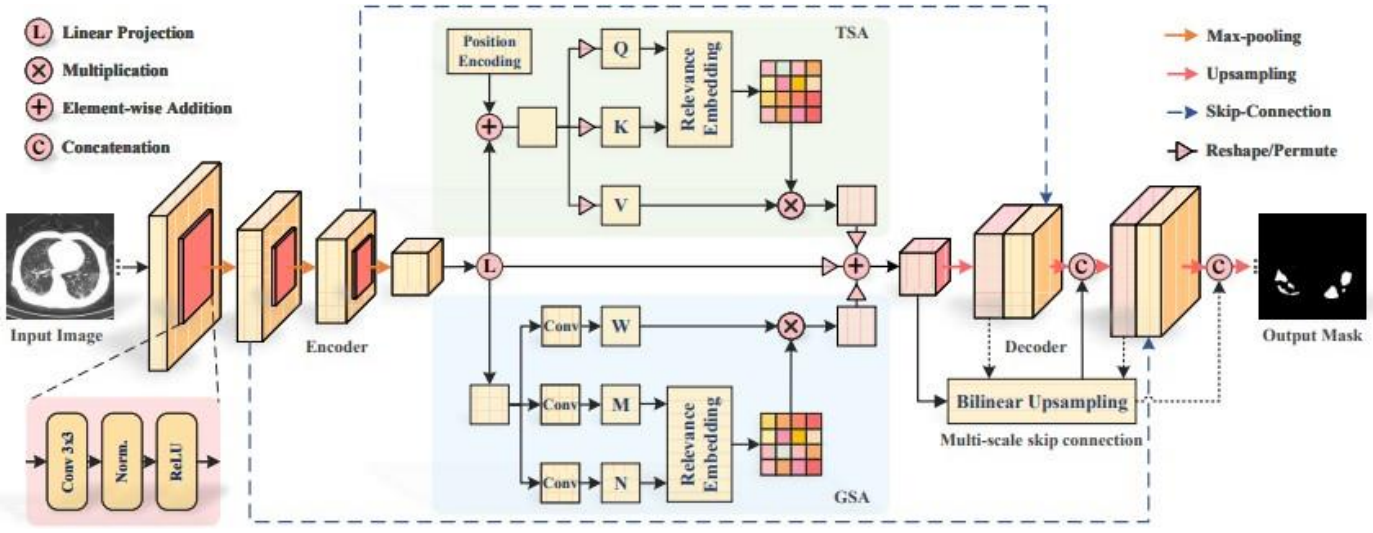
Multi-level Attention-guided U-Net with Transformer for Medical Image Segmentation

Intro

This paper introduces TransAttUnet, a novel medical image segmentation method that integrates Transformer-based attention mechanisms into the U-Net architecture. This integration enhances the model's ability to capture long-range dependencies and global contextual information.

TransAttUnet employs a self-aware attention (SAA) module, combining Transformer Self Attention (TSA) and Global Spatial Attention (GSA), to learn non-local interactions among encoder features. Multi-scale skip connections between decoder blocks aggregate features at different semantic scales, improving the representation of multi-scale context information.

These innovations mitigate detail loss from convolutional layers and sampling operations, enhancing segmentation quality. Experiments on various medical image datasets demonstrate that TransAttUnet consistently outperforms state-of-the-art methods.



How TransAttUnet Works:

1. **Encoder-Decoder Architecture**: TransAttUnet follows a traditional U-Net encoder-decoder structure, where the encoder extracts hierarchical features from the input image, and the decoder generates segmentation masks from these features.

2. **Multi-level Guided Attention Blocks**: These blocks incorporate Transformer-based self-attention mechanisms to capture long-range interactions and global spatial relationships. This attention mechanism helps focus on relevant features while suppressing irrelevant information, enhancing the model's discriminative power.

3. **Multi-scale Skip Connections**: Multi-scale skip connections between the encoder and decoder stages facilitate the propagation of contextual features from different semantic scales. This allows the decoder to refine segmentation masks with fine-grained details, improving accuracy and precision.

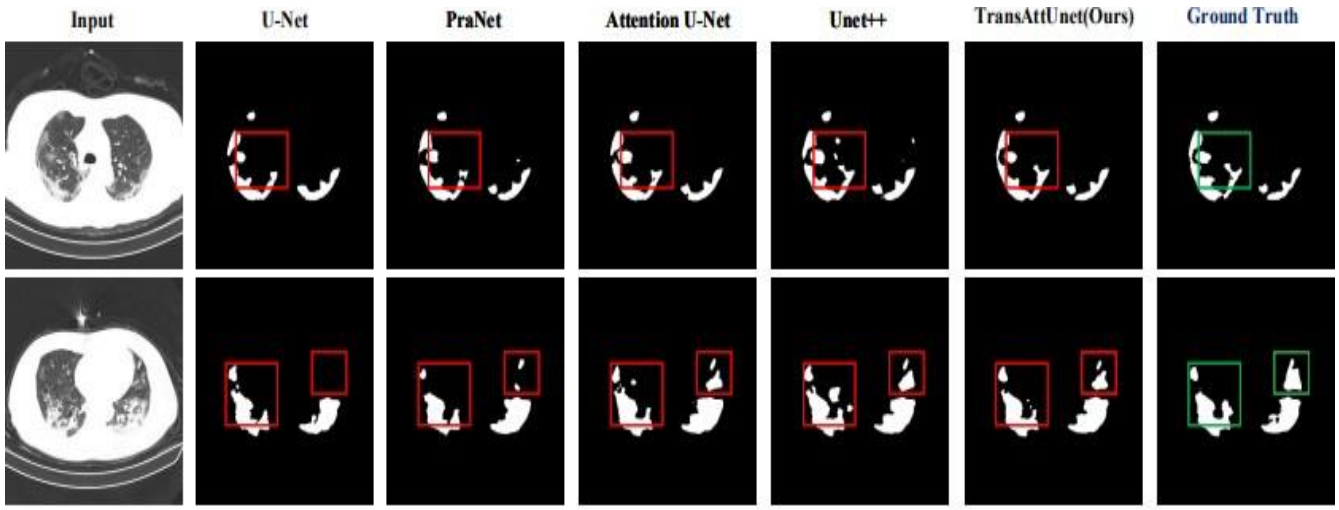
4. **Global Contextual Information**: Transformer-based attention mechanisms enable the model to capture global contextual information and long-range dependencies by allowing each encoder feature to deal with all other features. This captures intricate spatial relationships and semantic dependencies, essential for handling complex structures in medical images.

5. **Decoder Stage**: The decoder uses aggregated features from the encoder, combined with learned global contextual information, to generate segmentation masks. Attention mechanisms guide the decoder to focus on relevant features and suppress noise, resulting in more accurate and reliable segmentation.

TransAttUnet leverages the strengths of both U-Net and Transformer architectures, effectively capturing long-range dependencies and global contextual information. This

results in superior performance, particularly in challenging scenarios like biomedical imaging and brain image segmentation.

Results and comparisons:



(c) Quantitative results for pneumonia lesion segmentation.

TABLE I

COMPARISONS WITH THE STATE-OF-THE-ART BASELINES ON THE ISIC-2018 DATASET. ALL RESULTS WERE ANALYSED IN PERCENTAGE (%) TERMS. RESULTS OF THE MODEL WITH "*" ARE REIMPLEMENTED BY THE RELEASED SOURCE CODES. THE "-" DENOTES THE CORRESPONDING RESULT IS NOT PROVIDED. FOR EACH COLUMN, THE BEST AND SECOND BEST RESULTS ARE HIGHLIGHTED IN RED AND BLUE, RESPECTIVELY.

Method	Year	DICE	IoU	ACC	REC	PRE
U-Net [6]	2015	67.40	54.90	-	70.80	-
Attention U-Net [10]	2018	66.50	56.60	-	71.70	-
R2U-Net [38]	2018	67.90	58.10	-	79.20	-
Att R2UNet [38]	2018	69.10	59.20	-	72.60	-
ResUNet* [42]	2019	79.15	70.15	92.28	82.43	84.77
Channel-UNet* [8]	2019	84.82	75.92	94.10	94.01	81.04
BCDU-Net [43]	2019	85.10	-	-	7850	-
FANet [39]	2021	87.31	80.23	-	86.50	92.35
PraNet* [40]	2021	87.46	80.23	95.37	91.28	87.59
DoubleU-Net [36]	2020	89.62	82.12	-	87.80	94.59
Swin-Unet* [44]	2021	89.72	82.90	-	90.32	92.04
SegFormer* [45]	2021	90.24	83.60	-	91.12	92.10
MCTrans [30]	2021	90.35	-	-	-	-
TransAttUnet_C	-	89.25	81.46	95.06	89.90	91.59
TransAttUnet_D	-	90.14	83.04	96.14	90.42	92.17
TransAttUnet_R	-	90.74	83.80	96.38	90.93	92.42

Reference for this part: Multi-level Attention-guided U-Net with Transformer for MedicalImage Segmentation

Data Explanation

Based on the project requirements, we identified the need for two distinct types of datasets:

1. **Segmentation Dataset**
2. **Classification Dataset**

We have compiled and prepared both types of datasets, and we will detail their specifications below.

Image Classification vs Image Segmentation

Image Classification and Image Segmentation are two fundamental tasks in computer vision, each with distinct objectives and applications. Here is a concise explanation of each:

Image Classification

- **Objective:** To categorize an entire image into one of several predefined classes or labels.
- **Example:** Determining whether an image contains a dog or not, or classifying handwritten digits from 0 to 9.
- **Process:** The model analyzes the entire image and assigns it a single label based on its content.

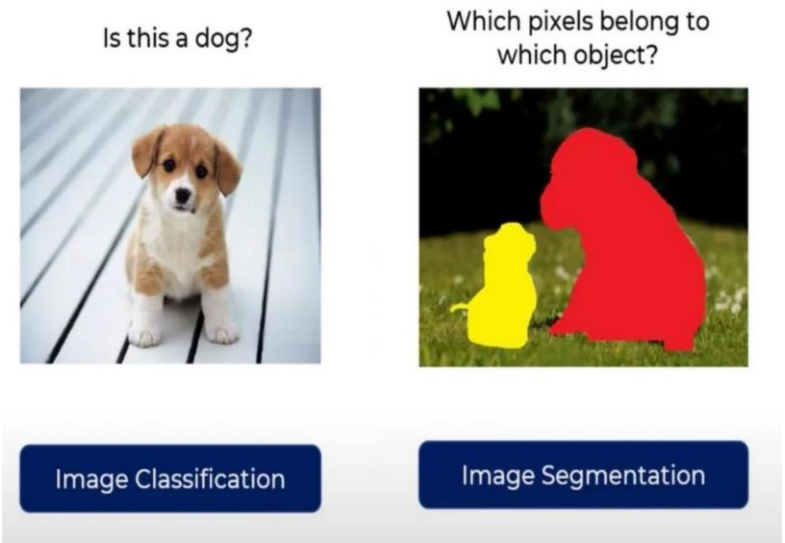


Image Segmentation

- **Objective:** To assign a label to every pixel in an image, grouping pixels with similar characteristics together.
- **Example:** In an image containing both a cat and a dog, segmentation algorithms label the pixels to identify and separate the regions corresponding to the cat and the dog.
- **Process:** The model examines each pixel and determines which object or region it belongs to, creating a detailed map of the image's content.

Segmentation Data

pre-operative multimodal MRI scans of glioblastoma (GBM/HGG) and lower grade glioma (LGG), with pathologically confirmed diagnosis and available OS, are provided as the training, validation, and testing data for Brain Tumor Segmentation (BraTS) challenge.

Brats Data context

BraTS utilizes multi-institutional pre-operative MRI scans and primarily focuses on the segmentation of intrinsically heterogeneous (in appearance, shape, and histology) brain tumors, namely gliomas. Furthermore, to pinpoint the clinical relevance of this segmentation task, BraTS also focuses on the prediction of patient overall survival , and the distinction between pseudo progression and true tumor recurrence, via integrative analyses of radiomic features and machine learning algorithms. Finally, BraTS intends to evaluate the algorithmic uncertainty in tumor segmentation.

Data Description

All BraTS multimodal scans are available as NIfTI files (.nii.gz) which consists of:

a) native (**T1**)

b) post-contrast T1-weighted (**T1Gd**)

c) T2-weighted (**T2**)

d) T2 Fluid Attenuated Inversion Recovery (**T2-FLAIR**) volumes

and were acquired with different clinical protocols and various scanners from multiple (n=19) institutions.

Classification Data

Classification data contains MRI data. The images are already split into Training and Testing folders.

The dataset contains images of human brain MRI images which are classified into 4 classes which are: Glioma, Meningioma, No tumor, Pituitary.

MRI scan channels

I. Native T1-Weighted Imaging (T1): Native T1-weighted magnetic resonance imaging (MRI) scans provide high-resolution anatomical details, commonly utilized to visualize brain structures. The term "native" indicates that these images are acquired without the administration of any contrast agents or special processing techniques.

II. Post-Contrast T1-Weighted Imaging (T1Gd): Post-contrast T1-weighted MRI scans are obtained following the administration of a gadolinium-based contrast agent. The gadolinium contrast enhances the visualization of specific brain features, such as blood vessels and regions with a disrupted blood-brain barrier, which are indicative of pathologies such as tumors or inflammatory processes.

III. T2-Weighted Imaging (T2): T2-weighted MRI scans are sensitive to variations in water content within tissues, making them particularly useful for detecting abnormalities such as edema, cysts, and lesions with high water content. These images typically exhibit high contrast between gray and white matter, revealing details that may not be as apparent on T1-weighted images.

IV. T2 Fluid Attenuated Inversion Recovery (T2-FLAIR) Volumes: T2-FLAIR MRI scans represent a specialized form of T2-weighted imaging that suppresses the signal from cerebrospinal fluid (CSF). This suppression enhances the visualization of pathological features while minimizing the signal from CSF, making T2-FLAIR particularly effective for detecting abnormalities in proximity to CSF spaces, such as periventricular lesions and white matter hyperintensities.

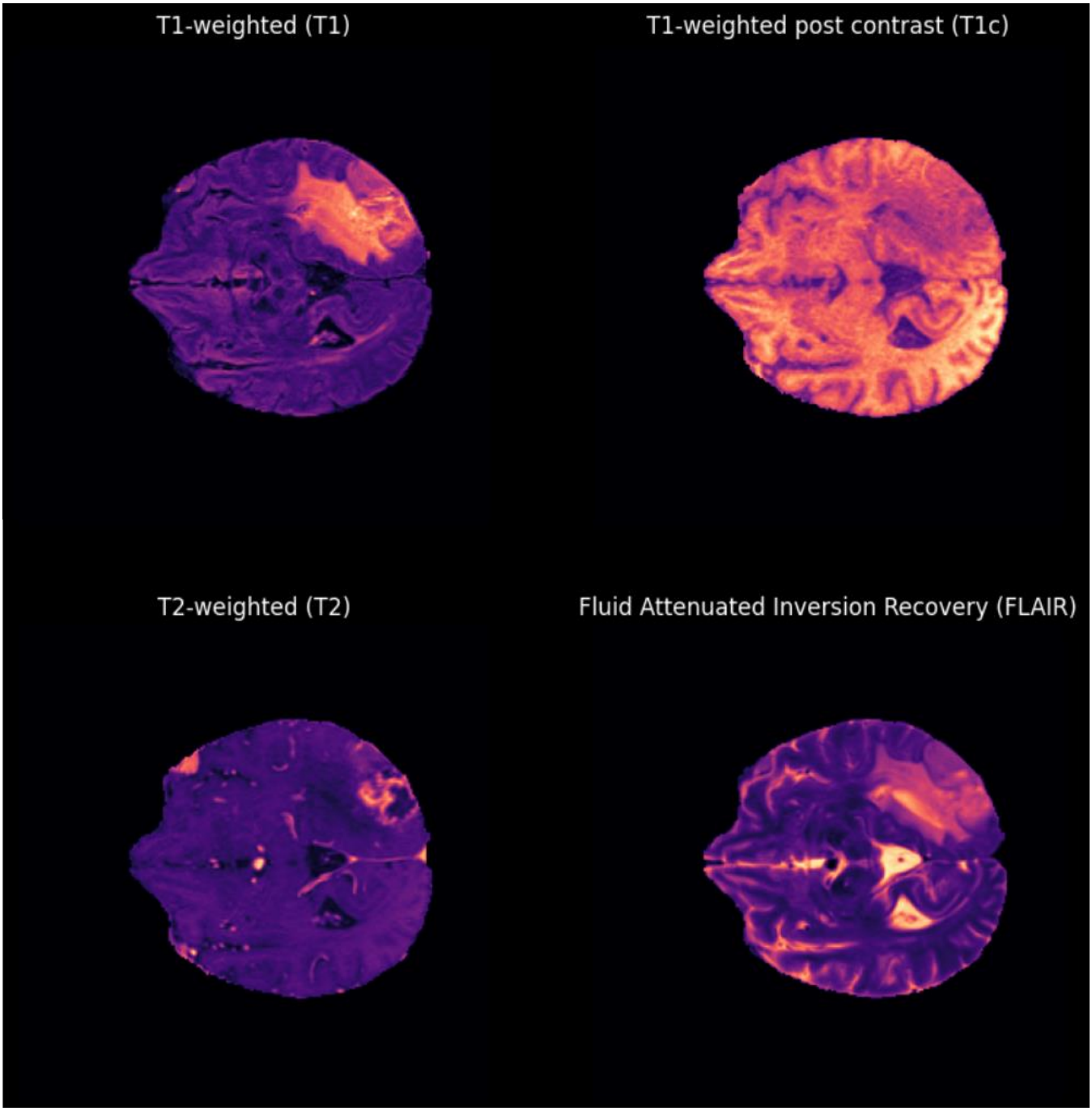
Mask channels

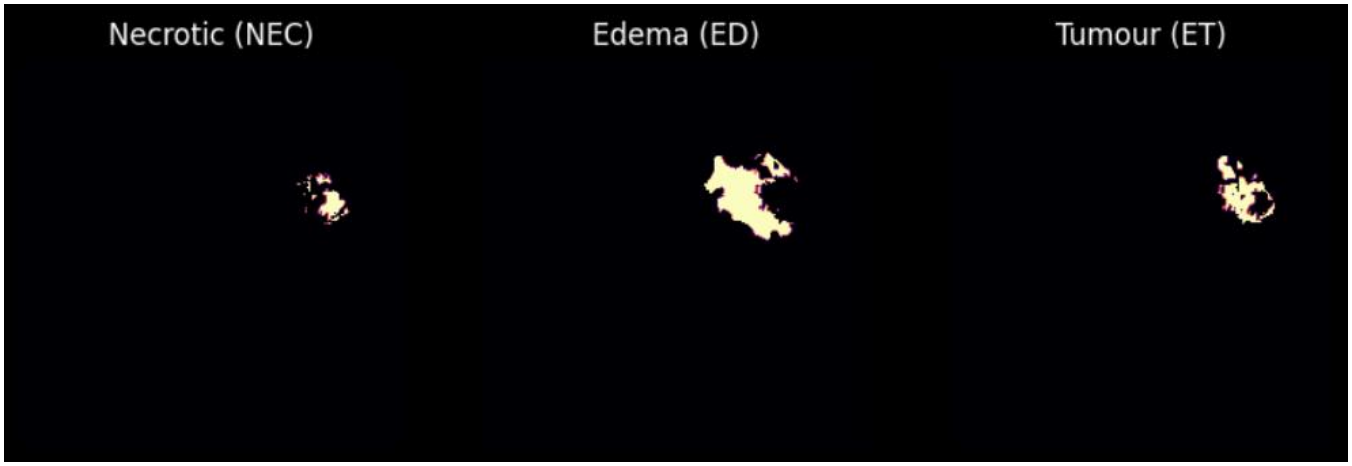
I. **Necrotic and Non-Enhancing Tumor Core (NCR/NET)**: This masks out the necrotic (dead) part of the tumor, which doesn't enhance with contrast agent, and the non-enhancing tumor core.

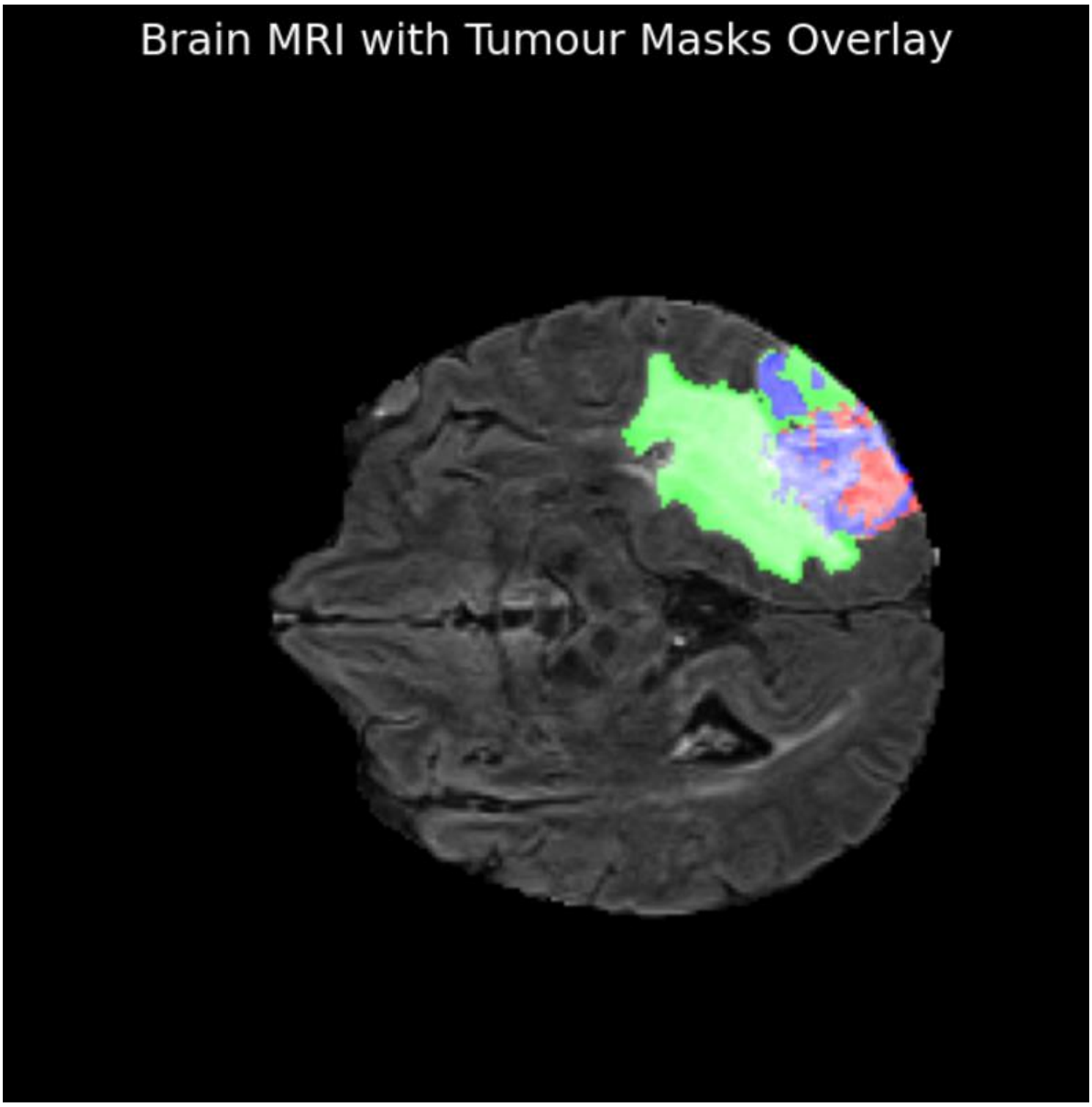
II. **Edema (ED)**: This channel masks out the edema, the swelling or accumulation of fluid around the tumor.

III. **Enhancing Tumor (ET)**: This masks out the enhancing tumor, which is the region of the tumor that shows uptake of contrast material and is often considered the most aggressive part of the tumor.

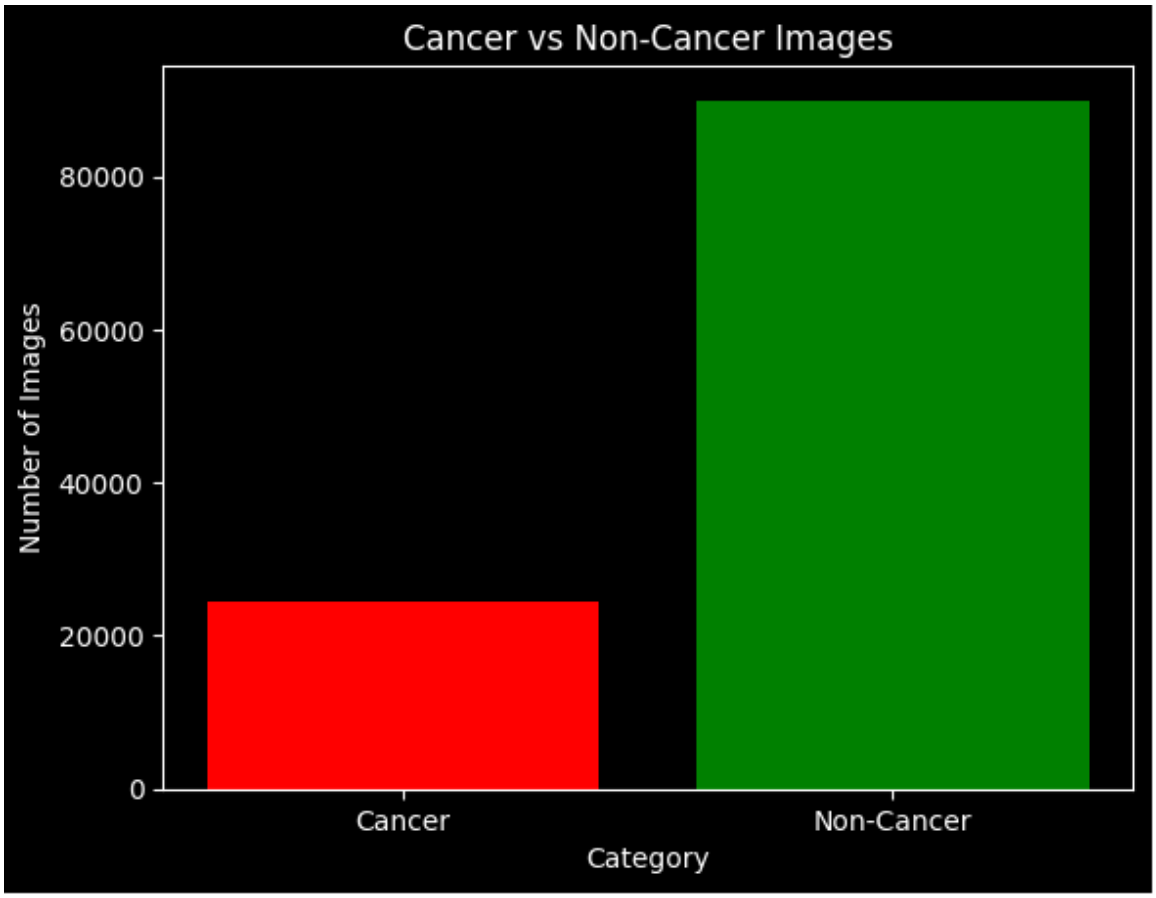
Data Visualization




Necrotic (NEC)
Edema (ED)
Tumour (ET)


Brain MRI with Tumour Masks Overlay

Cancer Vs Non-cancer



Feature Extractor

The feature extractor performs several key preprocessing steps:

1. **Image Resizing:** The image is resized to the dimensions expected by the SegFormer model. This ensures consistency and that the image dimensions match the input requirements of the model.
2. **Normalization:** Depending on the model requirements, the image pixel values might be normalized. This means adjusting the pixel values to a standard range (e.g., 0 to 1 or -1 to 1), which helps in stabilizing and speeding up the training and inference processes. In this case, normalization is skipped (do_rescale=False).
3. **Label Reduction:** If do_reduce_labels is set to True, the feature extractor reduces the number of labels in the segmentation map. This is helpful for models trained on datasets with many fine-grained categories where you might only care about more broad categories.
4. **Converting to Tensors:** The images are converted into PyTorch tensors. This format is required for processing the images with PyTorch-based models.
5. **Transferring to Device:** The tensors are moved to the specified device (CPU or GPU). This is crucial for utilizing hardware acceleration during model inference.

Important Note: This Feature Extractor is the premade function from hugging face directly to perform the basic preprocessing mentioned in the SegFormer paper.

References for this part:

1) Brain Tumor Segmentation(BraTS2020) (kaggle.com)

2) BRaTS 2021 Task 1 Dataset (kaggle.com)

3) Brain Tumor Classification (MRI) (kaggle.com)

4) Brain Tumor MRI Dataset (kaggle.com)

5) First Classification Data Description

6) SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, Ping Luo

7) RSNA-ASNR-MICCAI Brain Tumor Segmentation (BraTS)

Training GPU specifications

Parameter	Kaggle Kernel
GPU	Nvidia P100
GPU Memory	16GB
GPU Memory Clock	1.32GHz
Performance	9.3 TFLOPS
Support Mixed Precision	No
GPU Release Year	2016
No. CPU Cores	2
Available RAM	12GB
Disk Space	5GB

Note: All training was done on Kaggle.com.

Trials

SegFormer

SegFormer-Unet

MobileVitV2

Swin-Unet

SegFormer

Simple and Efficient Design for Semantic Segmentation with Transformers

Introduction

SegFormer introduces a novel approach for semantic segmentation that leverages the capabilities of Transformers while maintaining a simple and efficient design. The goal is to combine the strengths of Transformers in capturing long-range dependencies with the efficiency of lightweight models.

Key Concepts

1. Transformers:
    - Originally designed for language processing[2], transformers are now being adapted for image tasks. They can model long-range dependencies, meaning they understand relationships between pixels that are far apart in the image[3].

2. Multi-Headed Attention[3]:
    - A core component of transformers, multi-headed attention allows the model to focus on different parts of the image simultaneously, capturing various features and relationships at multiple levels of abstraction[3].

3. Hierarchical Transformer Encoder:
    - **Multi-Scale Feature Extraction**: SegFormer uses a hierarchically structured Transformer encoder that processes input images at multiple scales. This design captures wide global features and narrow local features at the same time.[4][1].
    - **Efficient Attention Mechanism[3]**: The encoder uses a new attention mechanism that is designed to be both computationally efficient and effective in capturing global context[3][1].

4. Lightweight MLP (Multi-Layer Perceptron) Decoder:
    - **Simple yet Effective**: The decoder in SegFormer is designed as a lightweight MLP that decodes the features extracted by the Transformer encoder. This approach combined with the decoder complexity is more efficient[1].

SegFormer Design Highlights

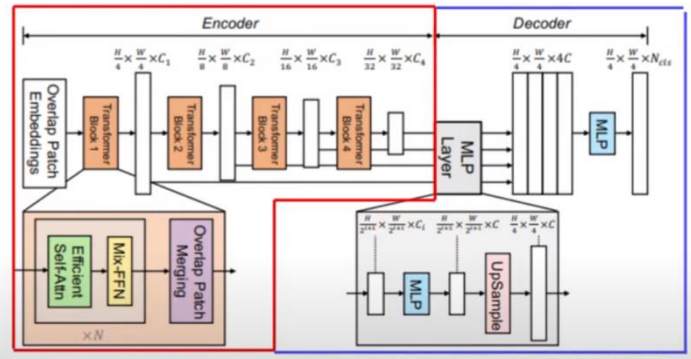


Figure 2: **The proposed SegFormer framework** consists of two main modules: A hierarchical Transformer encoder to extract coarse and fine features; and a lightweight All-MLP decoder to directly fuse these multi-level features and predict the semantic segmentation mask. "FFN" indicates feed-forward network.

Hierarchical Transformer Encoder:

The encoder captures features at multiple scales (or levels of detail). This means the model can understand fine details like edges and textures as well as broader regions like entire objects[4].

The encoder is structured in a way that it processes the image in stages, gradually building up a detailed understanding[4].

Multi-Headed Attention Mechanism:

This mechanism helps in learning the importance of different parts of the image. By using multiple heads, the model can attend to different parts of the image in parallel, making the feature extraction more robust and efficient[3].

Simple and Efficient Design:

Unlike many CNN-based methods that often require additional steps to refine the segmentation[5], SegFormer aims to keep the process straightforward. This simplicity helps in reducing computational requirements and improving processing speed[1].

Types of Classes:

In semantic segmentation, different types of classes refer to the categories into which pixels are classified. For example, in a street scene, the classes might include "road," "building," "tree," "car," and "pedestrian."

Performance

Benchmark Datasets: SegFormer has been tested on well-known datasets like ADE20K, Cityscapes, and COCO-Stuff. These datasets contain a variety of images and class labels, providing a rigorous test of the model's capabilities[1].

Results: The model achieves high accuracy and outperforms many existing methods while being more efficient in terms of computation and memory usage[1].

Used Pretrained Weights

MIT-B02

Performance in trial

Total Avg IoU: 0.6517

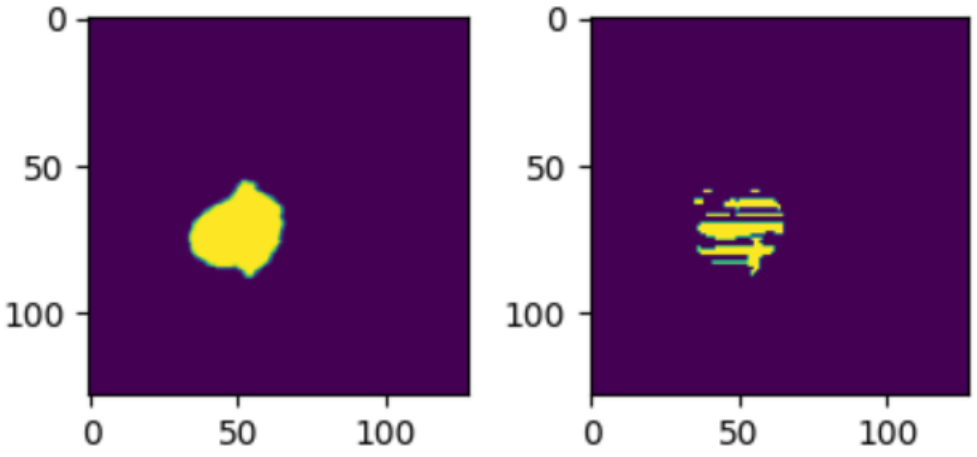
Total Avg Dice: 0.2710

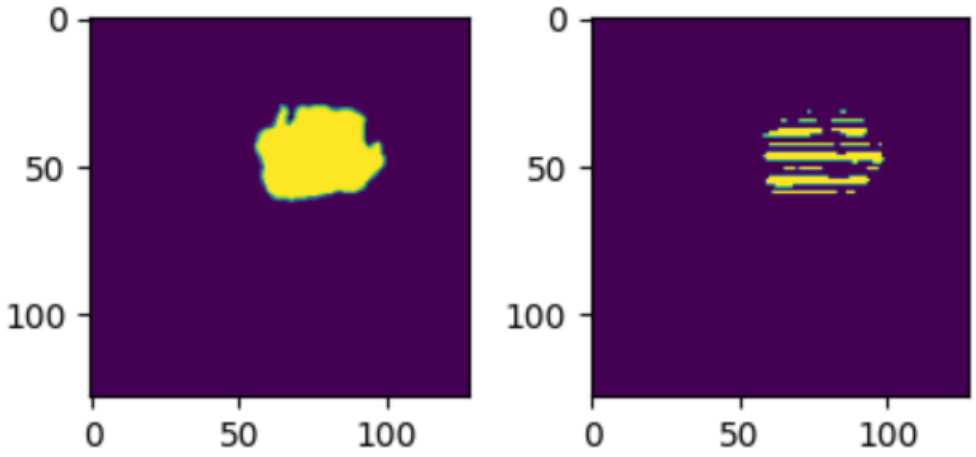
Total Avg Precision: 0.3598

Total Avg Recall: 0.0983

Total Avg F1: 0.1502

Example results





Data Preprocessing

I. Resize
II. Color Jitter
III. Combine t2 and flair channels

Advantages

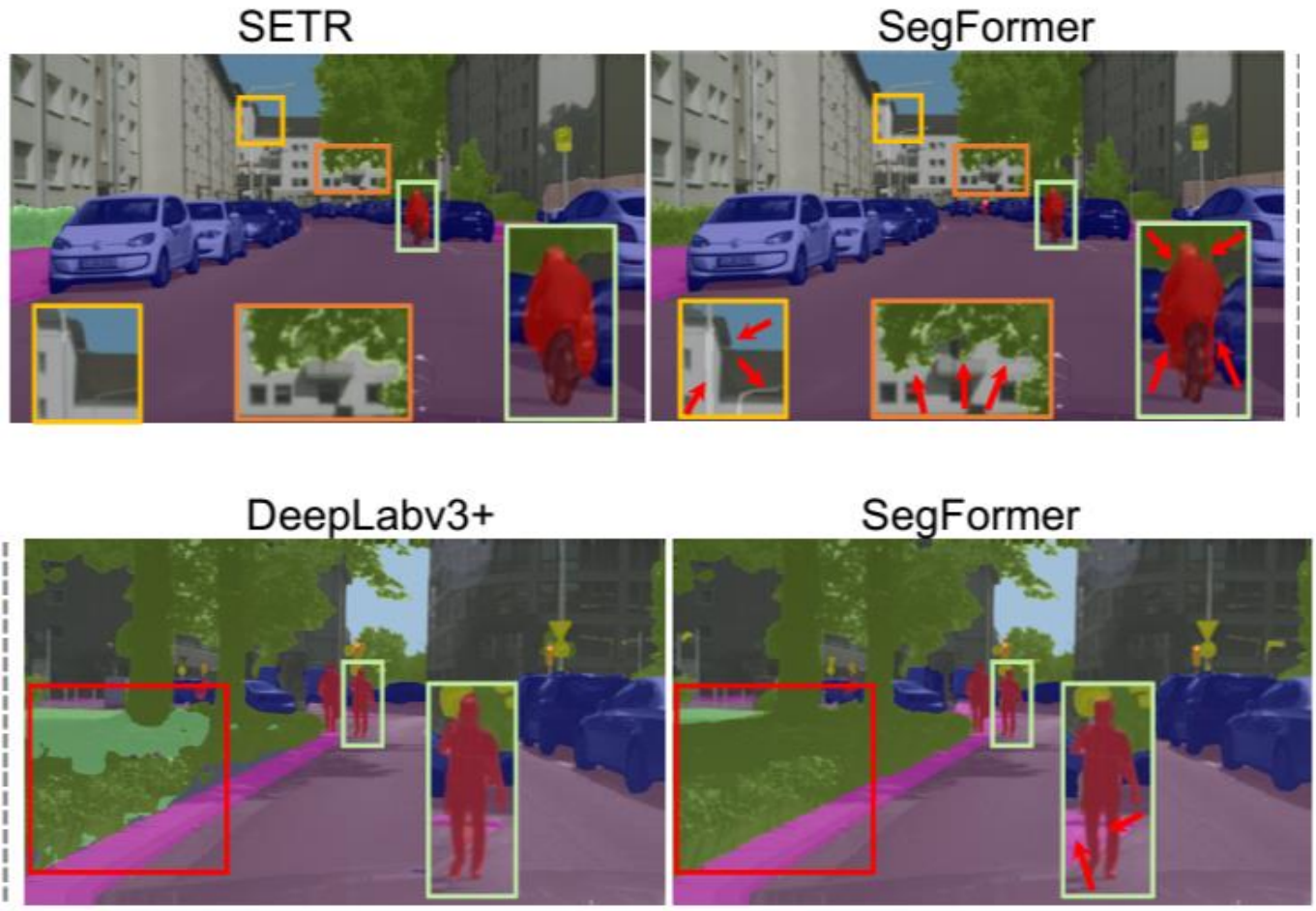
Simplicity: The design is straightforward and does not require complex post-processing steps[1].

Efficiency: Uses fewer computational resources compared to traditional CNN-based methods[1].

Flexibility: Can be applied to a wide range of images and segmentation tasks, making it versatile[1].

Applications

I. **Autonomous Driving**: The ability to efficiently and accurately segment road scenes makes SegFormer ideal for use in autonomous driving systems[1].
II. **Medical Imaging**: SegFormer's efficiency and performance make it suitable for medical image analysis, where accurate segmentation of anatomical structures is crucial[1].
III. **Robotics**: For robotics applications, where real-time processing is often required, SegFormer offers a good balance of accuracy and efficiency[1].



Conclusion

SegFormer introduces a novel approach to semantic segmentation by integrating transformers with a simple yet powerful design. The use of multi-headed attention and hierarchical encoding allows the model to effectively and efficiently capture the important features of an image. This results in a model that is both high-performing and practical for real-world applications.

In summary, SegFormer leverages the advanced capabilities of transformers to improve semantic segmentation, providing a balance of accuracy, efficiency, and simplicity. This makes it an exciting development in the field of computer vision.

References for this part:

1) SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, Ping Luo
2) Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. Advances in Neural Information Processing Systems, 30, 5998-6008.
3) Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. 2022. Transformers in Vision: A Survey. ACM Comput. Surv. 54, 10s, Article 200 (January 2022).
4) What Makes for Hierarchical Vision Transformer?: Yuxin Fang, Xinggang Wang, Rui Wu, Wenyu Liu
5) Zhao, X., Wang, L., Zhang, Y. et al. A review of convolutional neural networks in computer vision. Artif Intell Rev 57, 99 (2024).

SegFormer-Unet

Hybrid Transformer-CNN Architecture for Efficient and Accurate Semantic Segmentation

Key Concept

The SegFormer-UNet combined model is a type of neural network architecture used for image segmentation, which is the task of dividing an image into different regions or segments, often to identify objects or boundaries. This model leverages the strengths of two well-known architectures: SegFormer and UNet. The idea is to create a model that can accurately segment images while being efficient and scalable.

Architecture

A- SegFormer:

- **Backbone:** SegFormer uses a transformer-based backbone. Transformers are a type of model originally designed for natural language processing but have shown great promise in computer vision tasks. They excel at capturing global context, meaning they can understand the entire image at once rather than focusing on small parts.
- **Hierarchical Design:** SegFormer has a hierarchical structure, where the image is processed at multiple scales. This helps in capturing features at various levels of detail, from fine details to the overall structure.
- **Efficient Design:** Unlike other transformer-based models, SegFormer is designed to be computationally efficient, making it faster and less resource-intensive.

B- UNet:

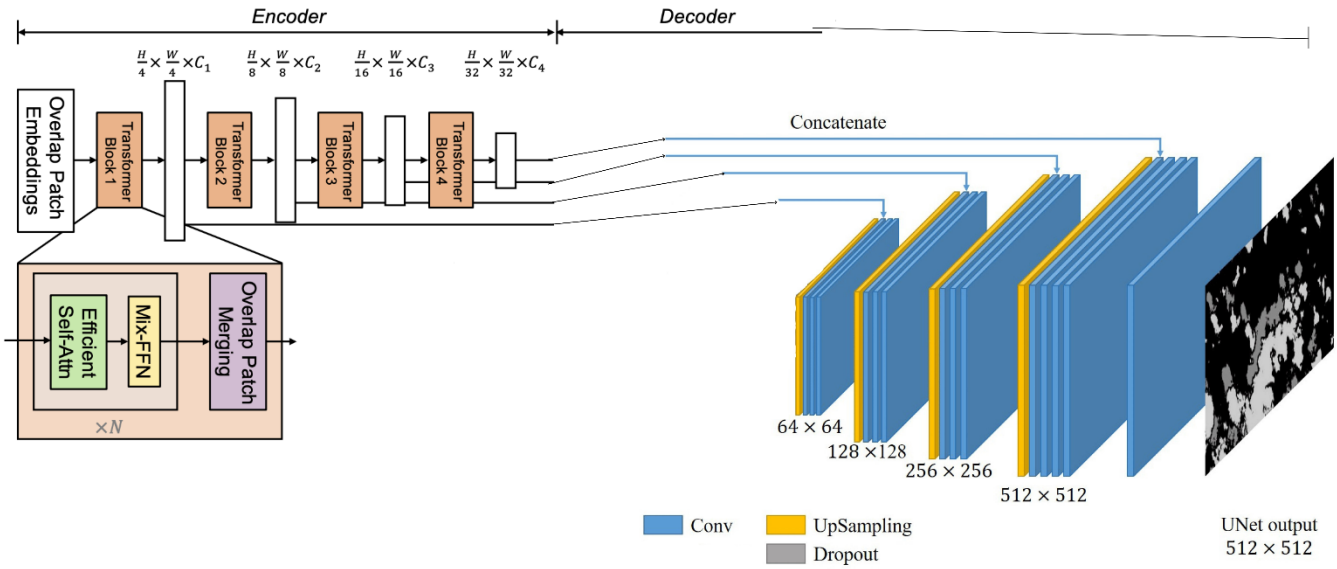
- **Encoder-Decoder Structure:** UNet is known for its encoder-decoder architecture. The encoder (contracting path) captures context by progressively down sampling the image, while the decoder (expanding path) reconstructs the image at the original resolution.
- **Skip Connections:** UNet includes skip connections that link the encoder and decoder at each level. These connections help the decoder to recover spatial information lost during down sampling, improving the segmentation accuracy.

Combined Architecture

The combined architecture integrates the strengths of SegFormer and UNet:

1. **SegFormer as Encoder:** The encoder part of the UNet is replaced with SegFormer. This means the initial image processing and feature extraction are done using SegFormer's transformer-based approach, which captures global context efficiently.
2. **UNet Decoder:** The decoder part remains the same as the traditional UNet architecture. It uses the hierarchical features extracted by SegFormer and reconstructs the segmented image, utilizing the skip connections to ensure precise localization.

Envisioned Combined Architecture:



Performance

The SegFormer-UNet combined model offers several advantages over each of them alone shown here:

U-Net [8]	38.30%
Seg-Net [9]	57.10%
Seg-Unet [15]	69.50%

(Note that this is the validation score on the CNUH dataset)[1]

Used Pretrained Weights

MIT-B02

Performance in trial

Total Avg IoU: 0.7833

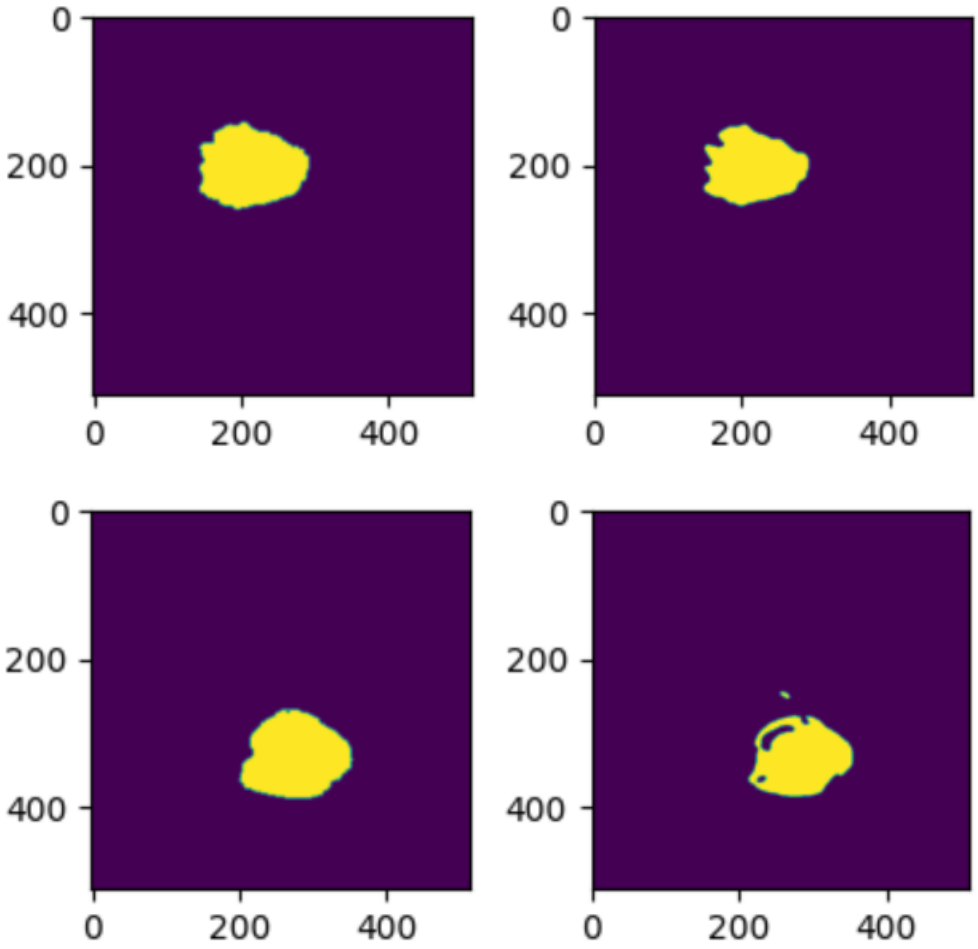
Total Avg Dice: 0.1618

Total Avg Precision: 0.3480

Total Avg Recall: 0.2209

Total Avg F1: 0.2599

Example results



Data Preprocessing

I. Resize
II. Color Jitter
III. Combine t2 and flair channels

Design Advantages

1. Enhanced Feature Representation:

    - **SegFormer** uses a hierarchical Transformer encoder to capture multi-scale features with strong global context awareness, providing better high-level representations.
    - **UNet** contributes with its skip connections that allow the combination of low-level and high-level features, which helps in retaining spatial details crucial for precise segmentation.

2. Flexibility and Adaptability:

    - **Versatility:** The combined model can be adapted to a wide range of segmentation tasks, from simple to complex scenarios, making it a versatile choice for different applications.
    - **Scalability:** It can efficiently scale with the complexity of the task, handling larger images and more intricate segmentation requirements without a proportional increase in computational cost.

3. Robustness:

    - **Resilience to Variations:** By leveraging the strengths of both SegFormer and UNet, the combined model can be more robust against variations in the data, such as changes in lighting, scale, and occlusions, which are common challenges in real-world segmentation tasks.

4. Multi-Scale Learning:

    - The architecture's ability to learn and fuse information at multiple scales results in better handling of objects of varying sizes within the same image, enhancing the overall segmentation performance.

Applications

- **Medical Imaging:** For tasks like tumor detection or organ segmentation in MRI or CT scans.
- **Autonomous Vehicles:** For segmenting road scenes to identify lanes, vehicles, pedestrians, etc.
- **Satellite Imaging:** For detailed land use and land cover classification, where capturing both large-scale patterns and small details is important. Land types include but are not limited to forests, urban areas, and bodies of water.
- **Agriculture:** For analyzing crop health or identifying different plant species.

Conclusion

The SegFormer-UNet combined model takes the best of both worlds: the global context understanding and efficiency of SegFormer and the precise localization and reconstruction capabilities of UNet. This results in a powerful, efficient, and flexible model that can handle a wide range of image segmentation tasks with high accuracy.

References for this part:

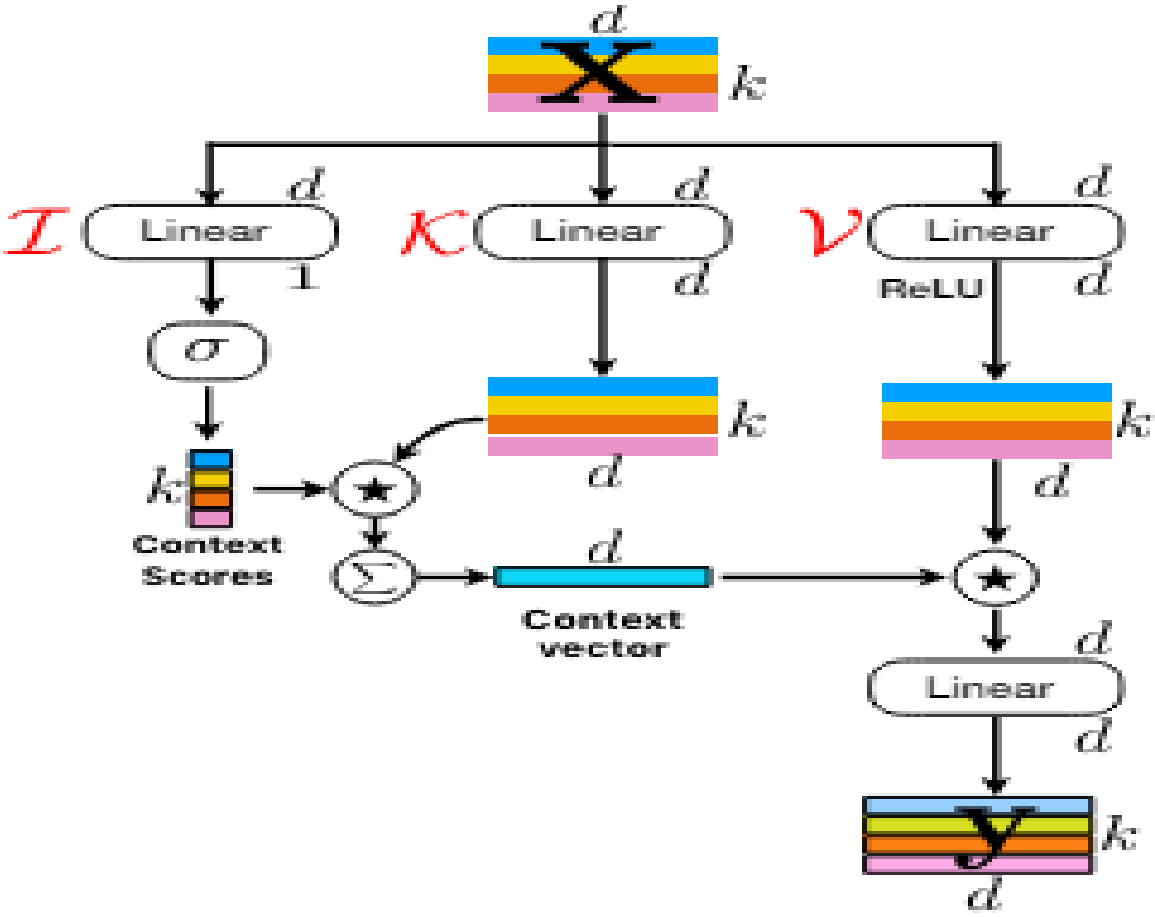
1) Do, N.-T.; Jung, S.-T.; Yang, H.-J.; Kim, S.-H. Multi-Level Seg-Unet Model with Global and Patch-Based X-ray Images for Knee Bone Tumor Detection. Diagnostics 2021, 11, 691.

MobileVitV2

Introduction

MobileViT V2: An Improved Version of MobileViT for “Efficient Image Recognition , Classification and Segmentation" focuses on enhancing the performance and efficiency of the original MobileViT model, which combines the strengths of convolutional neural networks (CNNs) and vision transformers (ViTs) for mobile and edge devices.

Architecture



Architecture details

- Hybrid Design:

- MobileViT V2 utilizes a hybrid architecture that integrates CNNs and ViTs. The CNN layers capture local features, while the ViT layers model long-range dependencies.

- Convolutional Layers:
  - The initial layers of the model are convolutional layers that process the input image to extract local patterns and features. These layers are optimized to be lightweight and efficient, suitable for mobile devices.

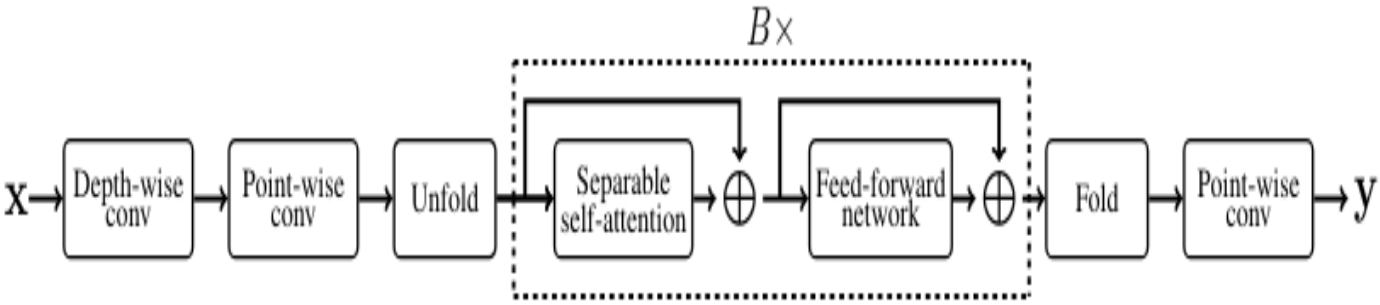
- Transformer Blocks:
  - The middle part of the architecture consists of transformer blocks. These blocks use multi-head self-attention mechanisms to capture global relationships between features. These blocks are designed to be more efficient compared to the original MobileViT by improving tokenization and reducing computational complexity.

- Efficient Tokenization:
  - In MobileViT V2, the tokenization process is enhanced to be more resource efficient. This involves converting image patches into tokens in a way that minimizes computational overhead while maintaining the ability to capture global context.

- Improved Attention Mechanisms:
  - The attention mechanisms within the transformers are optimized for better performance. This includes refining the self-attention calculations to be faster and more memory-efficient, which is crucial for deployment on resource-constrained devices.

- Fusion of Local and Global Features:
  - MobileViT V2 effectively combines local features from the CNN layers and global features from the transformer blocks. This fusion is carefully managed to ensure that the model benefits from both local and global information without significant increases in computational cost.

Detailed Architecture



Layer	Output size	Output stride	Repeat	Output channels
Image	256×256	1		
Conv-3×3, $\downarrow 2$ MV2	128×128	2	1 1	32α 64α
MV2, $\downarrow 2$ MV2	64×64	4	1 2	128α 128α
MV2, $\downarrow 2$ MobileViTv2 block (Fig. 6; $B = 2$)	32×32	8	1 1	256α $256 * \alpha$ ($d = 128\alpha$)
MV2, $\downarrow 2$ MobileViTv2 block (Fig. 6; $B = 4$)	16×16	16	1 1	384α 384α ($d = 192\alpha$)
MV2, $\downarrow 2$ MobileViTv2 block (Fig. 6; $B = 3$)	8×8	32	1 1	512α 512α ($d = 256\alpha$)
Global pool Linear	1×1	256	1	512α 1000

Used Pretrained Weights

ImageNet 1k

Performance in trial

Total Avg IoU: 0.6395

Total Avg Dice: 0.3239

Total Avg Precision: 0.2772

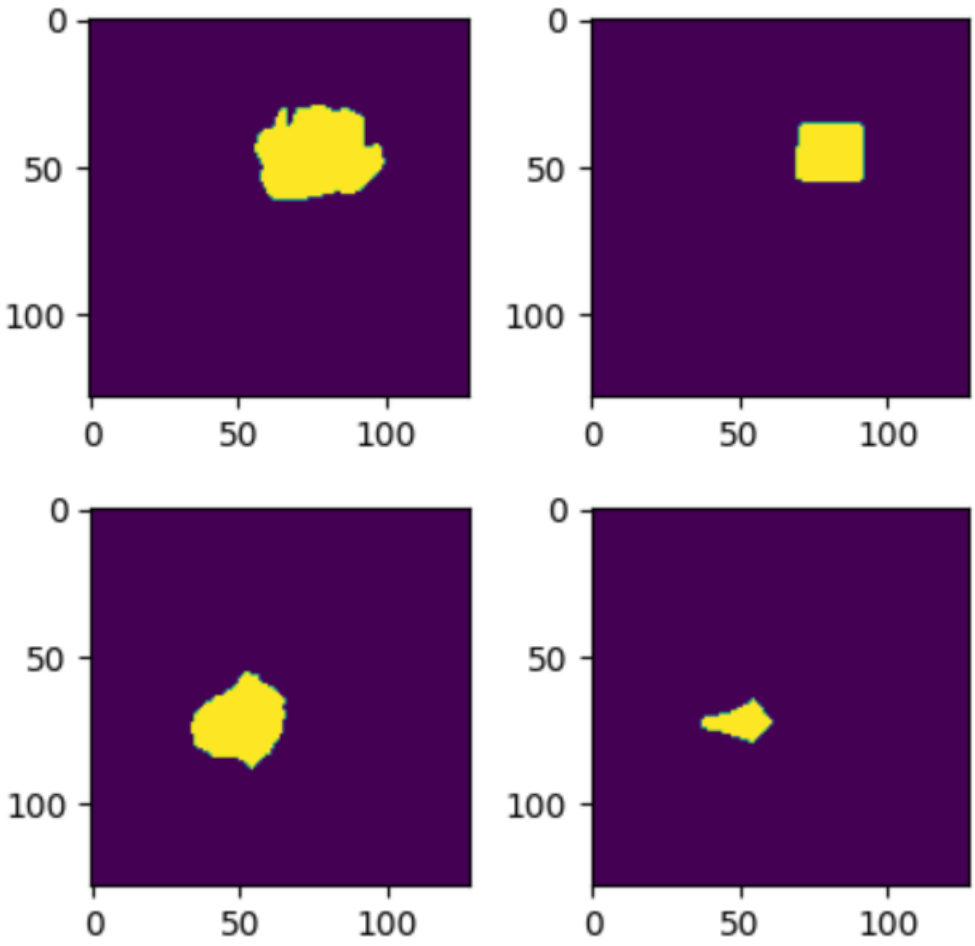
Total Avg Recall: 0.0654

Total Avg F1: 0.1018

Data Preprocessing

I. Resize
II. Combine t2 and flair channels

Example results



Reference for this part:

1) Separable Self-attention for Mobile Vision Transformers by : Sachin Mehta , Mohammad Rastegari

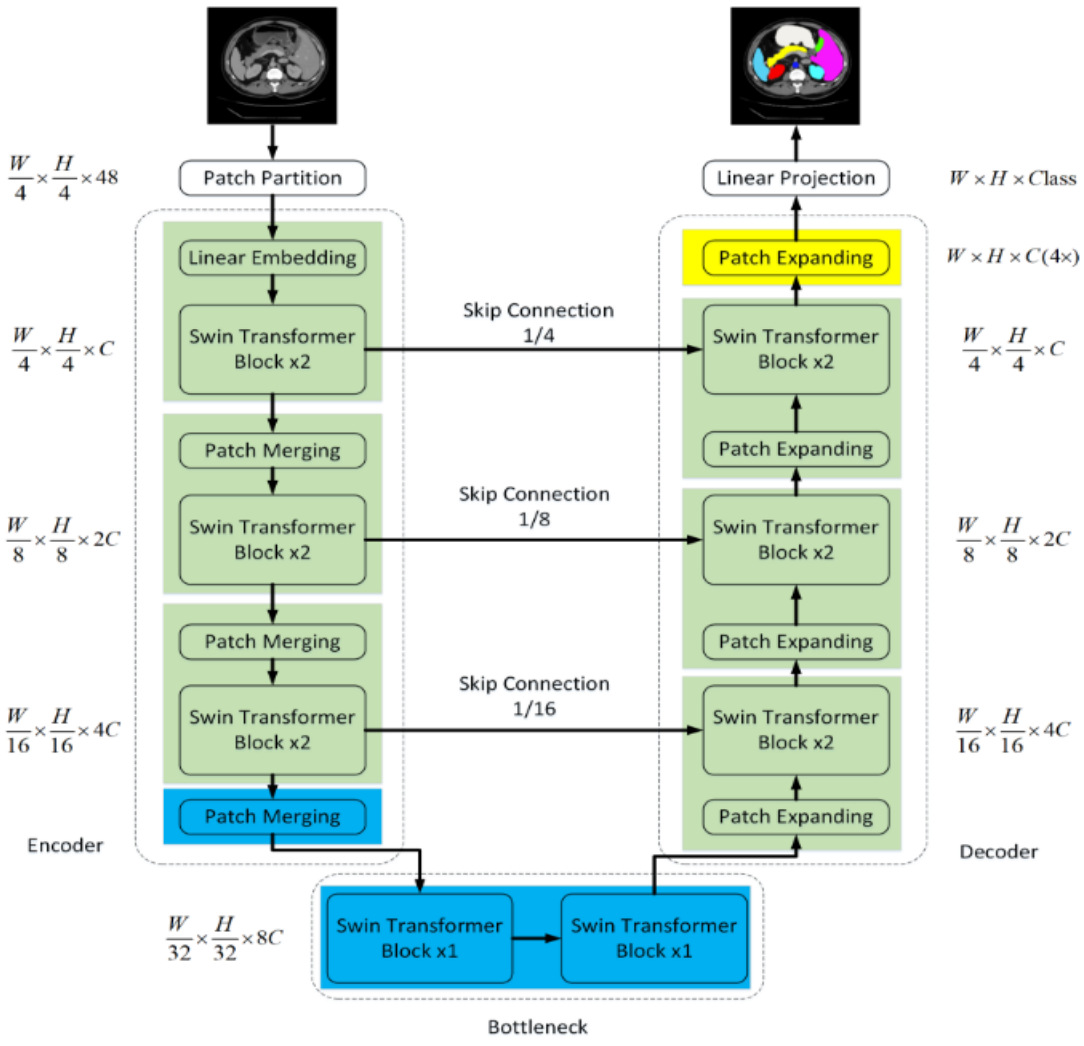
Swin-Unet

Efficient and Scalable Transformer for Medical Image Segmentation

Introduction

Swin-Unet is a transformer-based model designed for image segmentation tasks. It combines the hierarchical feature representation capabilities of Swin Transformers with the well-established U-Net architecture. The goal is to leverage the strengths of both approaches to achieve superior performance in segmenting medical images

Architecture



Components

- **Swin Transformer Backbone**:
    - **Hierarchical Representation**: The Swin Transformer processes images in a hierarchical manner, dividing them into non-overlapping patches. These patches are processed through multiple stages, each involving a Swin Transformer block that captures local and global context.
    - **Shifted Windowing**: Swin Transformer uses a shifted window mechanism to improve the efficiency and scalability of the model. This approach allows the model to capture cross-window interactions and long-range dependencies without excessive computational costs.

- **U-Net Structure**:
    - **Encoder-Decoder Framework**: Swin-Unet adopts the traditional U-Net architecture comprising an encoder and a decoder. The encoder captures features at multiple resolutions, while the decoder reconstructs the segmentation map from these features.
    - **Skip Connections**: Skip connections are used between corresponding layers of the encoder and decoder. These connections help in preserving spatial information and detailed features from the encoder, which are essential for accurate segmentation.

Detailed Workflow

1. **Input Image Processing**:
    - The input image is divided into non-overlapping patches. Each patch is then flattened and linearly embedded into tokens. These tokens serve as the input to the Swin Transformer blocks.
2. **Encoder Stage**:
    - **Patch Embedding**: The patches undergo a linear embedding to form the initial set of tokens.
    - **Swin Transformer Blocks**: The tokens pass through a series of Swin Transformer blocks. Each block consists of multi-head self-attention (MHSA) and multi-layer perceptron (MLP) layers, along with layer normalization and residual connections. The shifted window mechanism within these blocks enables efficient computation and long-range context capture.
    - **Patch Merging**: At the end of each stage, a patch merging layer reduces the number of tokens (down-sampling) while increasing their dimension, creating a hierarchical feature representation.

3. **Bottleneck Stage**:
    - The bottleneck stage further processes the encoded features using additional Swin Transformer blocks, capturing more complex patterns and deeper context information.
4. **Decoder Stage**:
    - **Patch Expanding**: The decoder uses patch expanding layers to up-sample the features, progressively increasing their resolution.
    - **Swin Transformer Blocks**: Similar to the encoder, the decoder includes Swin Transformer blocks to refine the features at each resolution level.
    - **Skip Connections**: The skip connections from the encoder are concatenated with the corresponding decoder features, helping to retain high-resolution information and detailed features.
5. **Output Segmentation Map**:
    - The final layer of the decoder generates the segmentation map, which is then reshaped to match the original input image dimensions. This map provides pixel-wise classification, delineating different regions of the image as required by the segmentation task.

Performance in trial

Average mIoU: 0.7470

Average Dice Loss: 0.2111

Average Precision: 0.3294

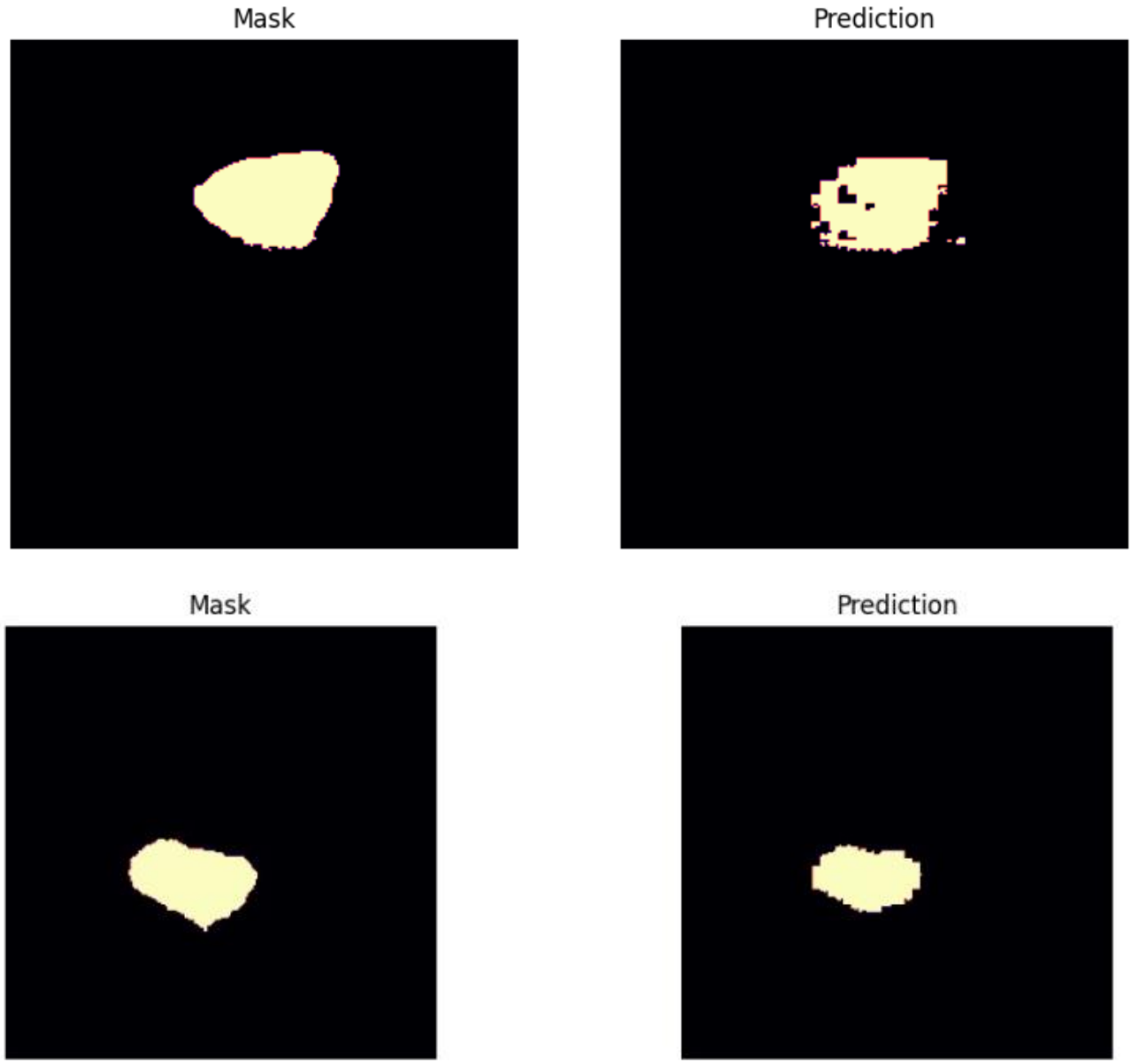
Average Recall: 0.3192

Average F1 Score: 0.3099

Data Preprocessing

I. Resize
II. Color Jitter
III. Filter out black images (images with pixels == 0)
IV. Horizontal , Vertical Flips
V. Combine t1 , t1ce channels (Since t1ce is the same as t1 but contrasted)
VI. Normalize pixel values by subtracting it's minimal value and dividing by its maximum plus a small constant to prevent division by zero

Example results



Reference for this part:

1) Swin-Unet: Unet-like Pure Transformer for Medical Image Segmentation Hu Cao 1 † , Yueyue Wang 2 † , Joy Chen 1 , Dongsheng Jiang 3 ∗ , Xiaopeng Zhang 3 ∗ , Qi Tian 3 ∗ , and Manning Wang 2

Architecture Backbone

Swin Transformer

Swin Transformer

Hierarchical Vision Transformer using Shifted Windows

Introduction

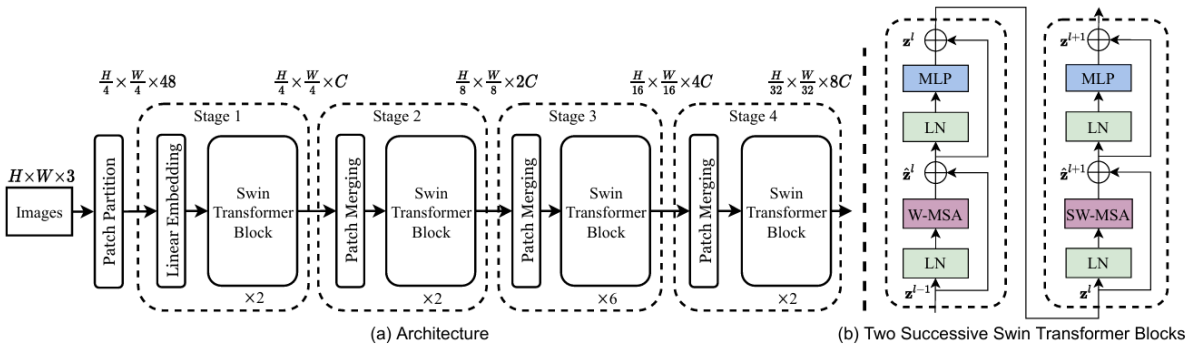
The introduction discusses the limitations of current Vision Transformers (ViTs) and convolutional neural networks (CNNs) in handling high-resolution images and dense prediction tasks. It introduces the Swin Transformer, a hierarchical transformer that computes self-attention within shifted windows. This design enables the model to achieve linear computational complexity with respect to image size, making it more efficient and scalable than previous methods. The Swin Transformer aims to combine the benefits of both CNNs and ViTs, achieving state-of-the-art performance on various vision tasks such as image classification, object detection, and semantic segmentation.

Related Work

This section reviews existing approaches in vision models, including CNNs, self-attention-based architectures, and hybrid models. It highlights the advantages and limitations of these methods. For instance, while ViTs have shown impressive results in image classification, their application to dense prediction tasks remains challenging due to quadratic complexity. The Swin Transformer is positioned as a solution that leverages local and global attention mechanisms to improve efficiency and effectiveness across different vision tasks.

Method

Overall Architecture



The Swin Transformer architecture starts by splitting an input RGB image into non-overlapping patches, treating each patch as a token. The dimensions and configurations used in this process are as follows:

1. **Patch Size and Embedding Dimension**: Each image is divided into patches of size 4x4 pixels. The initial embedding dimension of each patch is set to 96, meaning that each 4x4 patch (containing 48 raw pixel values, as each pixel has 3 RGB values) is projected into a 96-dimensional feature space through a linear embedding layer (ar5iv) (HugFace).
2. **Hierarchical Structure**: The architecture progresses through multiple stages:
    - **Stage 1**: The initial linear embedding layer processes the patches, keeping their number fixed while transforming their feature representation.
    - **Stage 2**: Patch merging layers reduce the number of tokens by concatenating features of neighboring patches, followed by a linear layer that projects these features into a 192-dimensional space (since each group of 2x2 neighboring patches results in a 4x96 dimensional input) (ar5iv).
    - **Stages 3 and 4**: This process repeats, further reducing the token count and increasing the feature dimension to 384 and then 768 in subsequent stages, maintaining a hierarchical feature representation suitable for various vision tasks.
3. **Attention Heads and Window Sizes**: The Swin Transformer uses a multi-head self-attention mechanism with varying numbers of attention heads across different stages: 3 heads in Stage 1, 6 in Stage 2, 12 in Stage 3, and 24 in Stage 4. The window size for local self-attention is fixed at 7x7 pixels (ar5iv) (HugFace).
4. **Configuration Parameters**: The architecture is fine-tuned using several parameters such as a multi-layer perceptron (MLP) ratio of 4.0, dropout probabilities, and specific activation functions (typically GELU) (ar5iv).

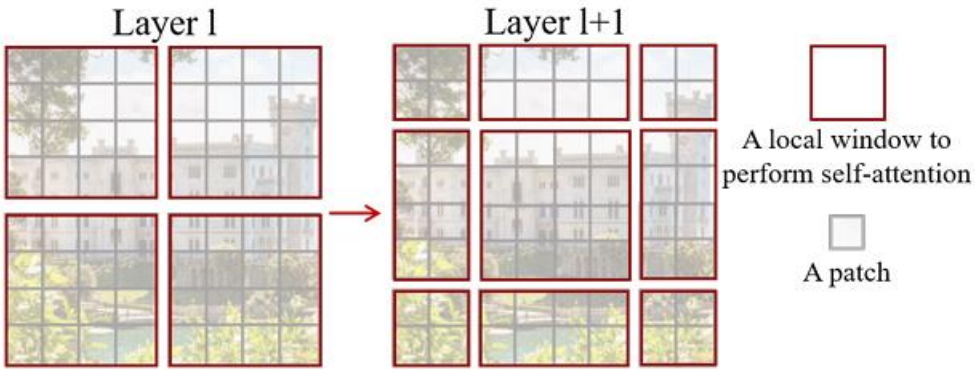
Modified Self-Attention with Shifted Windows

The Swin Transformer improves the standard self-attention mechanism by using **shifted windows**:

- **Standard Self-Attention**: Normally, self-attention is computed globally across the entire image, where each token attends to every other token. This is computationally expensive, especially for high-resolution images.
- **Shifted Windows**: The image is divided into small windows (e.g., 7x7 patches), and self-attention is computed within each window. This local self-attention reduces computational complexity. In alternating layers, the windows are shifted by a certain number of patches (e.g., shifting by half the window size). This shifting allows patches to attend to those in neighboring windows, enabling the model to capture global information through a series of local operations.

This hierarchical approach, combined with shifted windows, allows the Swin Transformer to manage computational complexity effectively while maintaining high-resolution feature maps. This makes it efficient for various vision tasks, including object detection and image segmentation, as it produces feature maps at multiple resolutions similar to conventional CNNs.

Shifted Window based Self-Attention



The core innovation of the Swin Transformer is its shifted window-based self-attention mechanism. Standard self-attention computes relationships between all token pairs globally, leading to high computational costs. The Swin Transformer, however, limits self-attention computation to local windows, reducing complexity to linear with respect to image size. To introduce cross-window connections and enhance modeling power, it alternates between two partitioning configurations in consecutive layers: regular and shifted windows. This approach maintains efficient computation while capturing global dependencies.

Experiments

The experimental results demonstrate the Swin Transformer's superior performance on three major vision tasks: image classification on ImageNet-1K, object detection on COCO, and semantic segmentation on ADE20K. The Swin Transformer outperforms previous state-of-the-art models in these tasks, showcasing its versatility and effectiveness. The experiments also include ablation studies that validate the importance of its design components, such as the hierarchical structure and shifted window mechanism.

Conclusion

The conclusion summarizes the contributions of the Swin Transformer, emphasizing its ability to handle high-resolution images and dense prediction tasks efficiently. The model's hierarchical design, combined with shifted window-based self-attention, provides a robust and scalable solution that advances the state-of-the-art in various vision applications. Future work includes exploring more applications and improving the model's efficiency further.

Our Model

BefUnet

Introduction and Motivation

1. Properties of Convolutional Neural Networks (CNNs)

a. **Local Translation Invariance**:
   - Once a CNN learns to recognize a feature in one part of an image, it can recognize the same feature in a different part of the image. This is due to the shared weights across the convolutional filters.

b. **Focus on Local Features**:
   - CNNs primarily focus on local features due to their convolutional nature. This can be a limitation when dealing with objects that require an understanding of the global context or when there is a need to capture long-range dependencies between different parts of the image.
   - Example: In medical imaging, where patient anatomy can vary greatly (large inter-patient variation), it is crucial to understand the entire context of the scan to accurately segment a tumor.

2. U-Net Limitations

- The U-Net architecture, with its skip connections, has proven versatile for segmentation tasks. However, the convolution layer's locality restriction limits its representational power in capturing shape and structural information crucial for medical image segmentation.

3. Vision Transformers

- Inspired by the recent success of transformers in Natural Language Processing (NLP), vision transformers have been developed to overcome CNN limitations in image recognition tasks. Vision transformers leverage multi-head self-attention (MSA) to establish long-range dependencies and capture global contexts.

  Challenges:

  a. **Data Demand**:
     - Transformers require large datasets to train effectively. They need to learn from many examples to understand the various patterns and relationships within the data. This can

be a limitation in fields like medical imaging, where large, annotated datasets are not always available.

b. **Quadratic Complexity**:

- The self-attention mechanism in transformers calculates the relationship between each pair of elements in the input data. As the size of the input grows, the number of pairwise comparisons grows quadratically. This means that for large inputs, such as high-resolution images, transformers can become computationally expensive and slow, requiring significant memory and processing power.

4. Vision Transformer (ViT) Models

- To address the limitations of CNN models, Vision Transformer (ViT) models utilize the MSA mechanism and achieve state-of-the-art (SOTA) performance compared to convolution-based methods.

5. Limitations of Transformer-only Models

- While transformer-based models can capture global feature representations at multiple levels, they may struggle to capture local features as effectively as CNNs. Local features such as edges, corners, and textures are important for segmentation tasks.

6. Hybrid CNN-Transformer Approaches

- Hybrid models, such as TransUnet and LeViT-Unet, have been proposed to combine the locality of CNNs with the long-range dependency of transformers. These models aim to encode both global and local features in medical image segmentation.

  Challenges:

  – These approaches face difficulties in effectively combining high-level and low-level features while maintaining feature consistency.
  – They may struggle to fully exploit the potential of multi-scale information produced by the hierarchical encoder, such as edge features.

7. Body and Edge Fusion Unet (BEFUnet)

- BEFUnet aims to achieve precise medical image segmentation by enhancing the fusion of edge and body information.

Architecture Abstract

Key Modules

Dual-Branch Encoder Module: The dual-branch encoder is designed to simultaneously extract edge and body information.

- **First Branch:** A lightweight CNNs branch with pixel-wise convolution.

- **Second Branch:** A hierarchical Transformer branch based on the SwinTransformer.

Double-Level Fusion (DLF): Responsible for merging features from the CNN and Swin Transformer levels while maintaining feature consistency. It utilizes a cross-attention mechanism to fuse information across scales, handling features of different sizes and balancing localization and semantic information (image context).

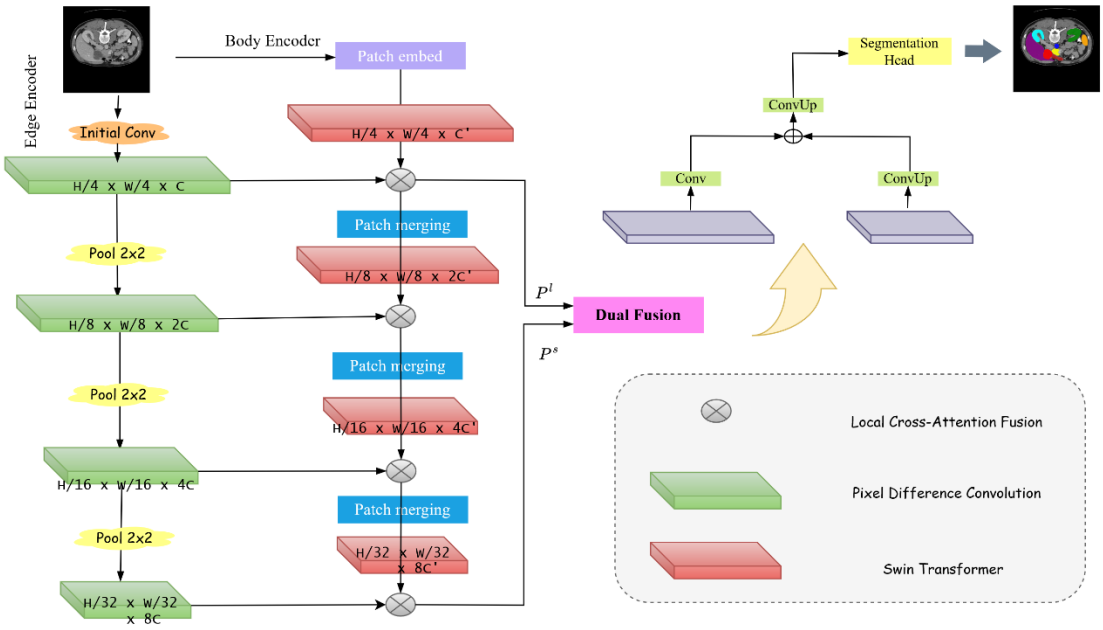
Local Cross Attention Feature Fusion:

- Increases segmentation accuracy by effectively fusing features that are closely located in position between the dual branches.
- Reduces computational complexity by only storing the features mentioned above, not all global features.

Combining Edge and Body Information

- By combining the edge local semantic information of CNN with the body contextual interactions of the transformer, the model enhances the integration of complementary features. This approach proves advantageous in handling irregular boundaries in medical image segmentation.

Architecture Details



Dual-Branch Encoder

Edge Encoder: Addresses the issues of inadequate edge information extraction by conventional medical segmentation networks through a dedicated edge detection branch.

PDC (Pixel Difference Convolution): Vanilla convolution calculates the weighted sum of pixel values within the convolution kernel, while PDC calculates the weighted sum of differences between pixel values within the kernel.

$$y = f(x;\theta) = \sum_{i=1}^{k} w_i \cdot x_i \quad \text{(vanilla convolution)} \quad (1)$$

$$y = f(\nabla x;\theta) = \sum_{(x_i,x_j)\in P} w_i \cdot (x_i - x_j) \quad \text{(PDC)} \quad (2)$$

PDC Block Components: PDC, ReLU layer, Convolution layer with kernel size of 1, Residual connection.

Process:

- **Stage 1:**
  - Reduce image size to ensure compatibility with the output size (H/4, W/4, C).
  - Apply PDC block.
  - Max Pooling (2x2).

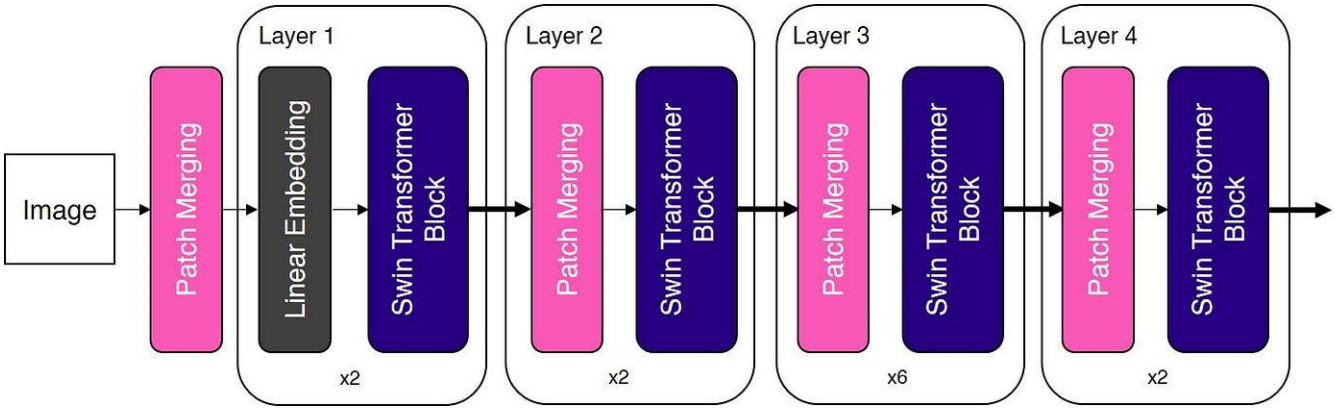
- **Stage 2:**
    - Reduce image size (H/8, W/8, 2C).
    - Apply PDC block.
    - Max Pooling (2x2).
- **Stage 3:**
    - Reduce image size (H/16, W/16, 4C).
    - Apply PDC block.
    - Max Pooling (2x2).
- **Stage 4:**
    - Reduce image size (H/32, W/32, 8C).
    - Apply PDC block.
    - Max Pooling (2x2).

After each stage, the resulting feature map is compared to the ground truth to further enhance edge extraction.

- **Loss Function:** Annotator-robust loss function.

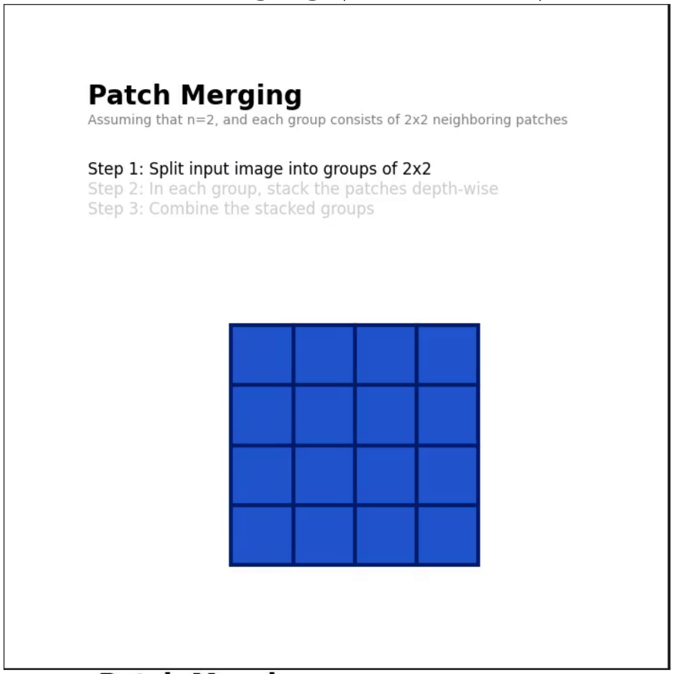
Body Encoder: - Utilizes the Swin Transformer architecture to capture global features.

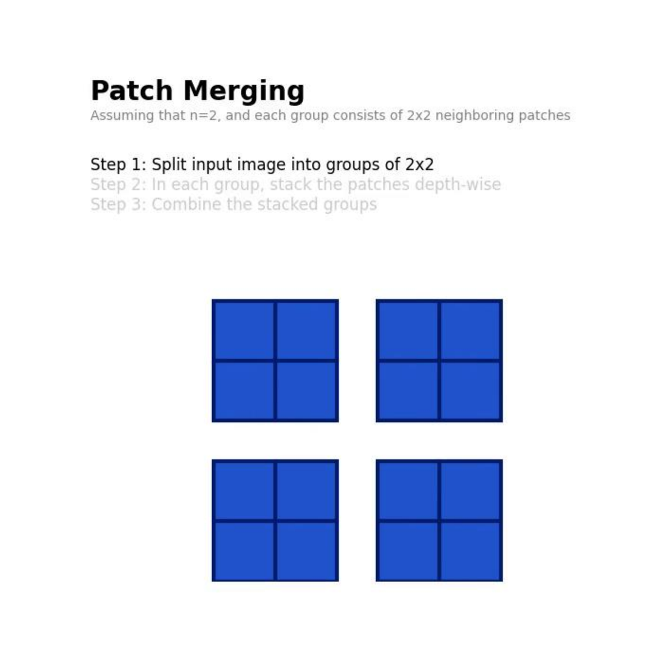
Swin Block Components: Patch Merging ,Window-MSA or shifted window-MSA, MLP, normalization layer applied after each MSA layer and MLP layer.

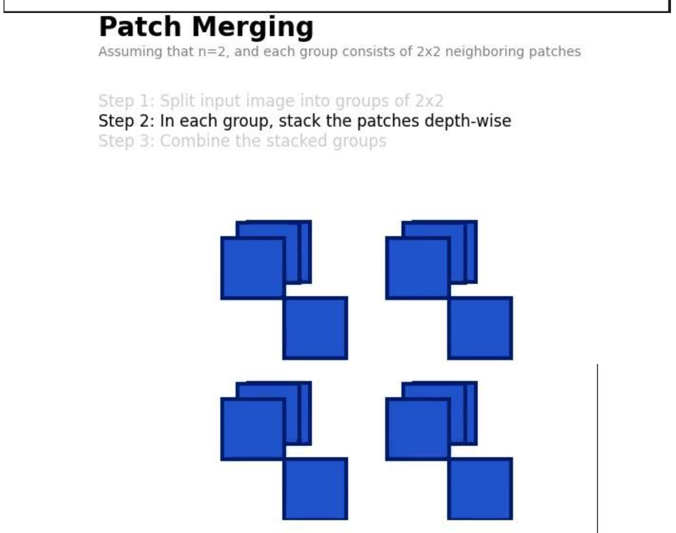


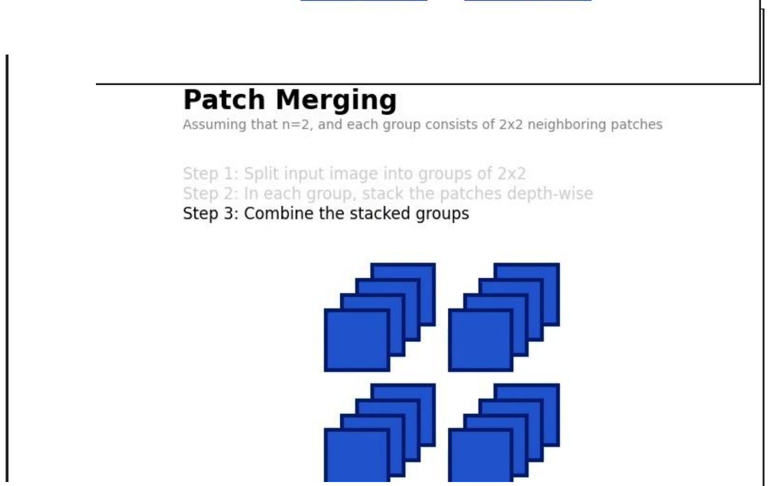
Explaining the Swin block main components for better understanding of BEFUnet's body Encoder

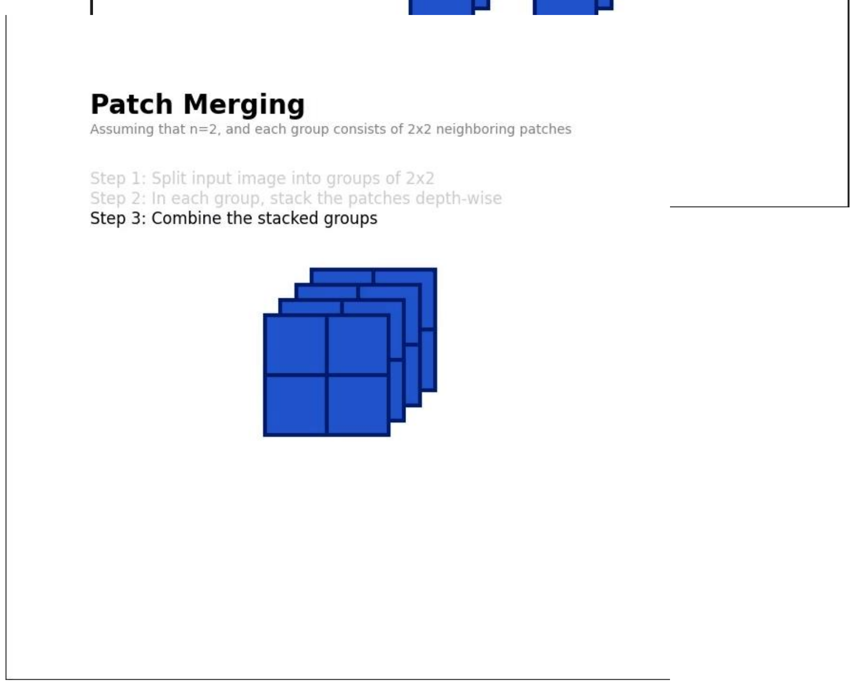
- Patch Merging (Illustration):





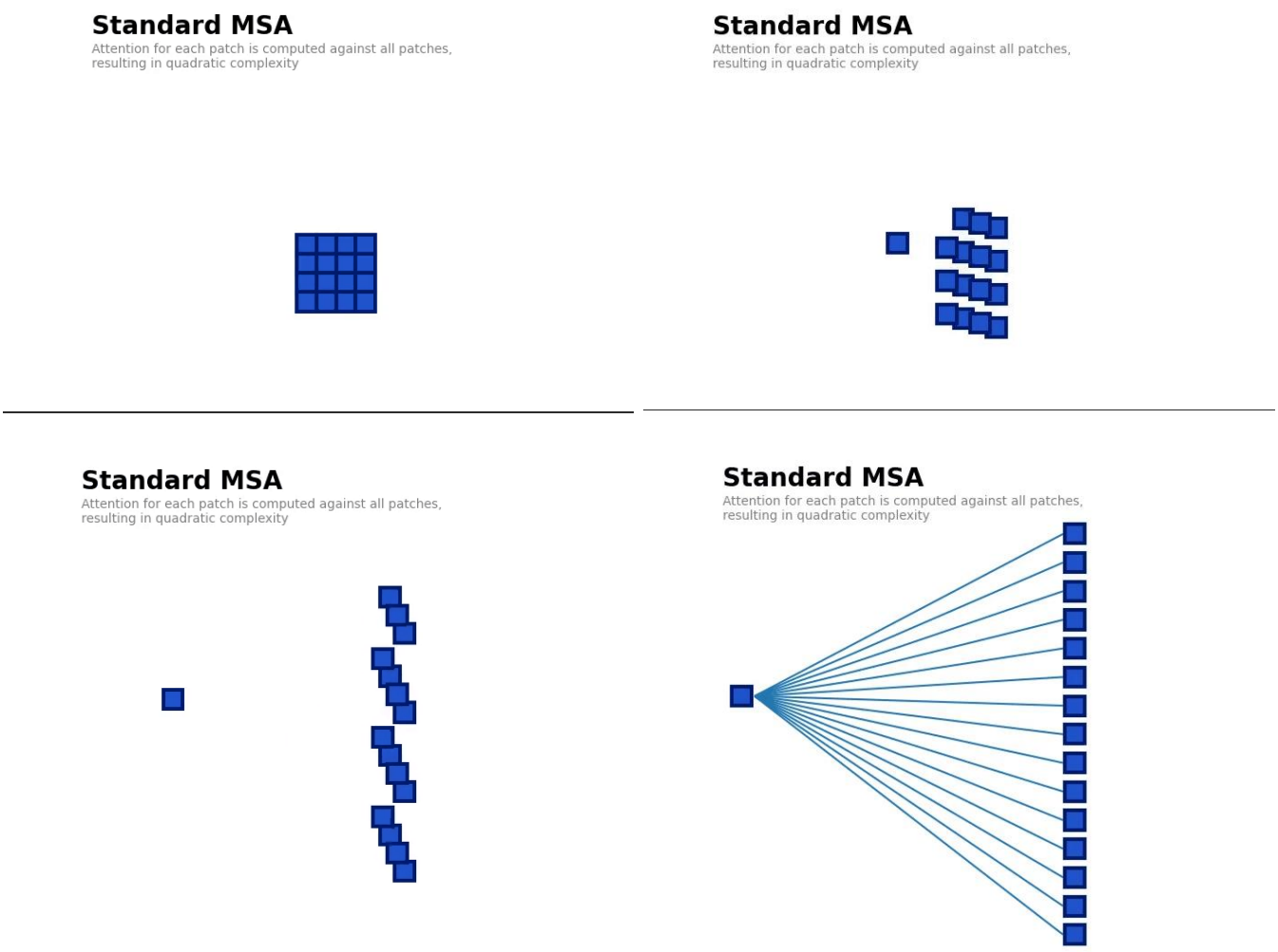




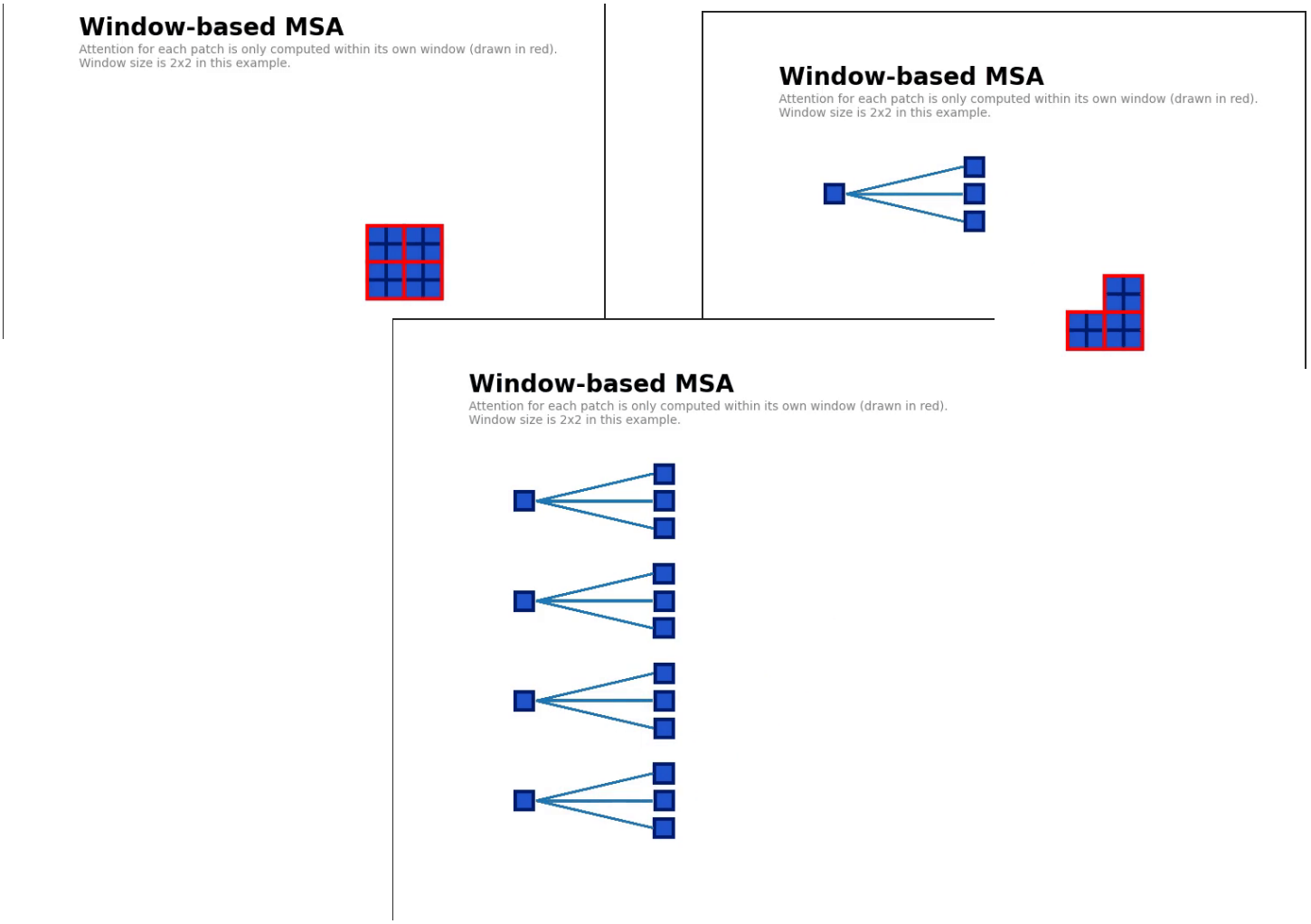


Standard Patch MSA as introduced in the original VIT vs Swin (W-MSA) & (SW-MSA):

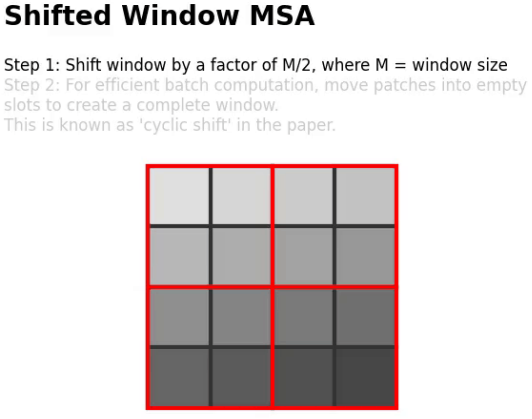
Standard Patch MSA: applies multi-head self-attention on for each patch against every patch which results in quadratic complexity as the image's resolution increases as illustrated in the figure



Windowed multi-head self-attention: it solves the problem with computation complexity by defining a window surrounding several patches and each patch will look to only its neighbors in the specified window

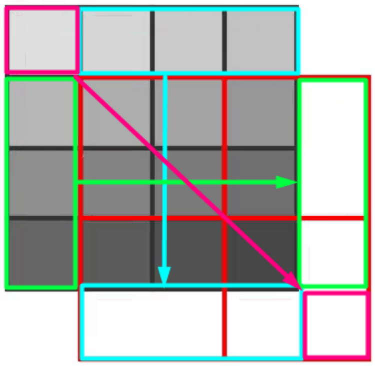


Shifted multi-head self-attention: although W-MSA solved the computation problem it restricted the model's attention field of view, solving this problem SW-MSA was introduced it shifts the windows towards the bottom right corner by a factor of M/2, where M is the window size, this shift results in 'orphaned' patches that do not belong to any window, solving this problem a Cyclic Shift technique was introduced which moves the 'orphaned' patches into windows with incomplete patches.



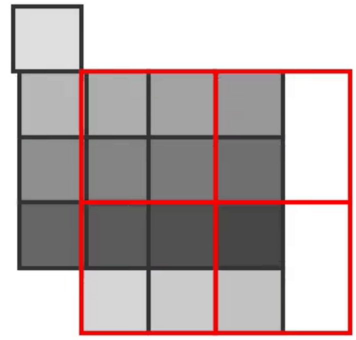
Shifted Window MSA

Step 1: Shift window by a factor of M/2, where M = window size
Step 2: For efficient batch computation, move patches into empty slots to create a complete window.
This is known as 'cyclic shift' in the paper.



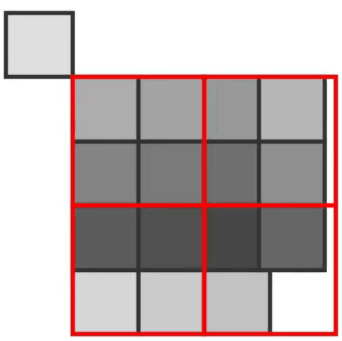
Shifted Window MSA

Step 1: Shift window by a factor of M/2, where M = window size
Step 2: For efficient batch computation, move patches into empty slots to create a complete window.
This is known as 'cyclic shift' in the paper.



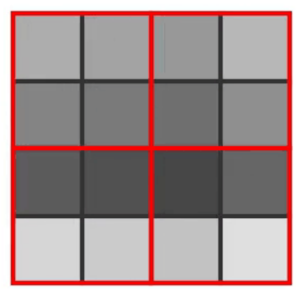
Shifted Window MSA

Step 1: Shift window by a factor of M/2, where M = window size
Step 2: For efficient batch computation, move patches into empty slots to create a complete window.
This is known as 'cyclic shift' in the paper.



Shifted Window MSA

Step 1: Shift window by a factor of M/2, where M = window size
Step 2: For efficient batch computation, move patches into empty slots to create a complete window.
This is known as 'cyclic shift' in the paper.



Continue Body Encoder

-Process:

- **Stage 1:**
  - Patch embedding.
  - Reduce image size to ensure compatibility with the output size (H/4, W/4, C).
  - Swin Block (W-MSA).
  - Patch merging (between Stage 1 edge and Stage 1 body).

- **Stage 2:**
  - Reduce image size (H/8, W/8, 2C).
  - Swin block (SW-MSA).
  - Patch merging (between Stage 2 edge and Stage 2 body).
- **Stage 3:**
  - Reduce image size (H/16, W/16, 4C).
  - Swin block (W-MSA).
  - Patch merging (between Stage 3 edge and Stage 3 body).
- **Stage 4:**
  - Reduce image size (H/32, W/32, 8C).
  - Swin block (SW-MSA).
  - Patch merging (between Stage 4 edge and Stage 4 body).

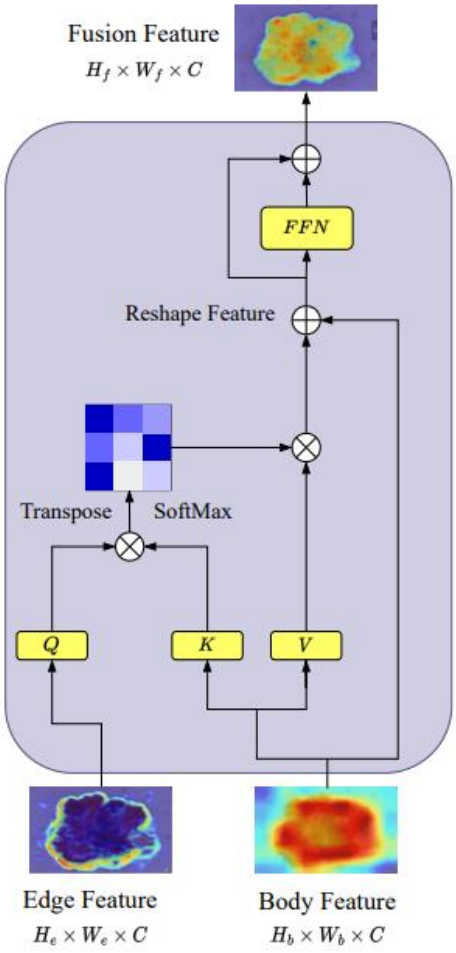
Loss Function: Binary cross entropy loss + weight * dice_loss.

LCAF Module

- Designed to fuse edge and body features accurately and efficiently by selectively performing cross-attention on features that are spatially close between the two images.
  - Input edge and body features are divided into local regions, projected into query, key, and value vectors, and attention scores are calculated through dot product. The resulting vectors are weighted and summed to obtain the fused feature representation.
  - Compared to conventional global cross-attention, LCAF significantly reduces computational complexity by performing selective local-cross attention.

Process:

1- **The input edge and body features are divided into local regions:** This division helps in focusing the attention mechanism on specific spatial areas rather than the entire feature map, which reduces computational complexity. Instead of computing a single attention function, the multi-head mechanism divides the input features into multiple subspaces and computes the attention function independently in each subspace.



captured in each subspace.

2- **Query, Key, and Value Vectors**: These local regions are then projected into query (Q), key (K), and value (V) vectors. This projection is essential for calculating attention scores, these vectors are split into several smaller groups. Each group represents a subspace. For example, if the feature dimension is 512 and there are 8 heads, each head operates in a subspace of dimension 64 (since 512 / 8 = 64).

3- **Attention Calculation**: Attention scores are calculated using the dot product of the query and key vectors, followed by a SoftMax operation to normalize these scores. (all heads' computations are independent from each other therefore they are parallel processed at the same time)

4- **Weighted Sum**: The resulting attention scores are used to weigh the value vectors. These weighted vectors are then summed to obtain the fused feature representation. Since all attention heads are independent, each subspace can focus on different parts or aspects of the input features, allowing the model to capture a richer and more diverse set of relationships.

5- **Concatenation of Results**: After computing attention scores and weighted sums within each subspace, the results from all subspaces (heads) are concatenated to form the final output, combining the diverse information

Double-Level Fusion (DLF)

- The DLF module effectively combines coarse(global context features) and fine-grained(local features) feature representations, increasing the strengths of both CNNs and transformers to produce a comprehensive and consistent feature map for medical image segmentation. DLF Solves the issue of merging data between the transformer and CNN while maintaining data consistency by taking the shallowest and the deepest levels as inputs. Shallow levels contain more precise localization information, while deeper levels carry more semantic information that is better suited for the decoder. (Middle layers are not added to the fusion to reduce computational power needed).

Process:

1- **Selection of levels**: takes in the shallowest and the deepest levels as inputs. Shallow levels contain more precise localization information, while deeper levels carry more semantic information

2- **Class Tokens**: For each selected level, a class token is generated through Global Average Pooling (GAP) and normalization. The class tokens summarize the information from the input features, ensuring that essential details from each level are retained.

3- **Embedding and Position Information**: The class tokens are concatenated with the respective level embeddings. Learnable position embeddings are added to each token to incorporate positional information, aiding the transformer encoders in understanding the spatial context of the features.

4- **Transformer Encoders**: The small and large levels are processed through a series of transformer encoders to compute global self-attention. The small level is followed by **S** transformer encoders, while the large level is followed by **L** transformer encoders. These encoders utilize the concatenated class tokens and position embeddings to enhance the fusion process.

Data Preprocessing

I. Random Flip
II. Random Rotation
III. Zoom
IV. Merge Channels using Max Function

Performance in Trial

Total Avg_MIoU : 0.805

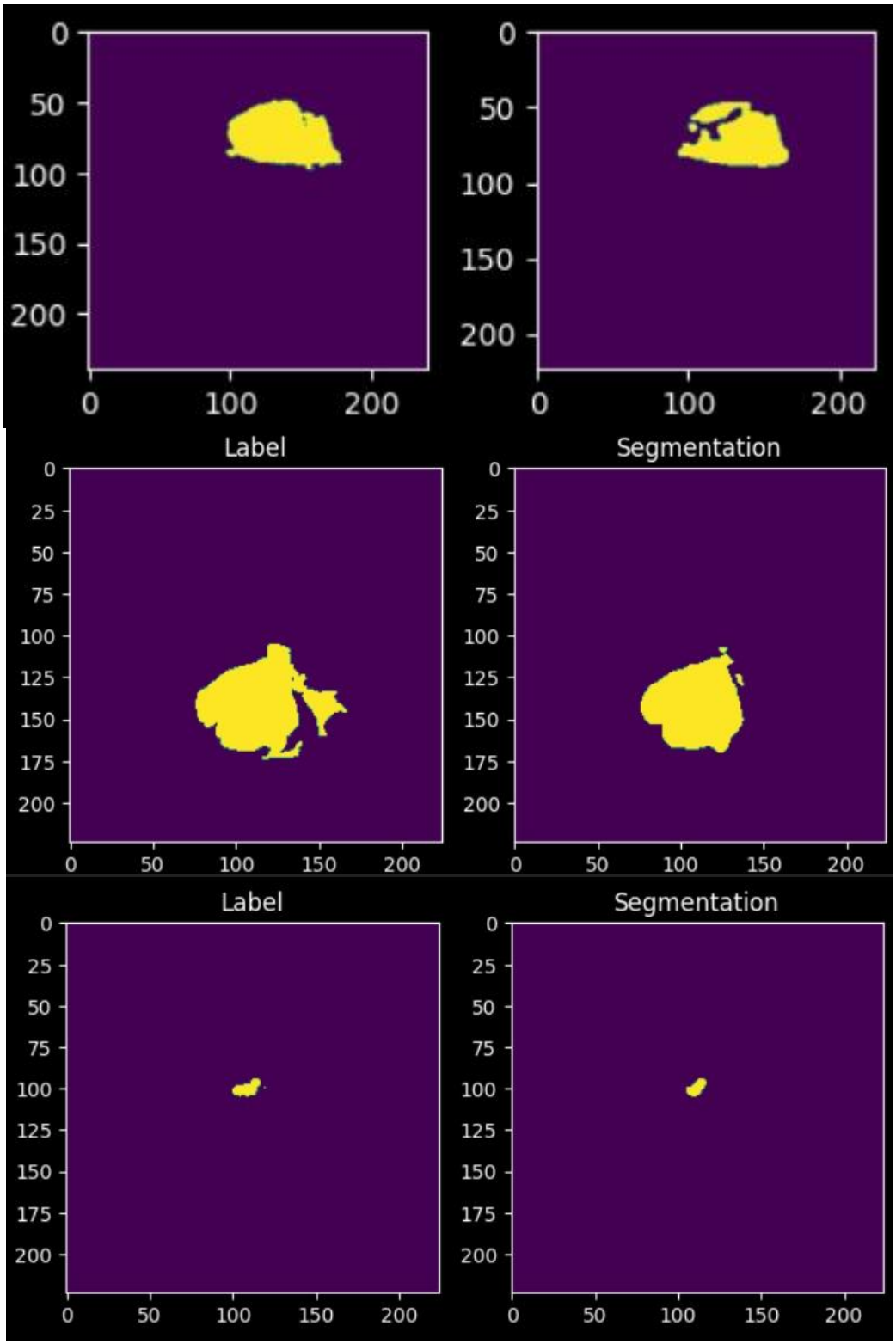
Total Avg_Dice_Coef : 0.154

Total Avg_Precision : 0.393

Total Avg_Recall : 0.335

Total Avg_F1_Score : 0.353

Example results



Glioma Grades Classification

Overview

After using Bef-Unet in segmentation, the segmented image is then introduced to a classifier.

Data processing

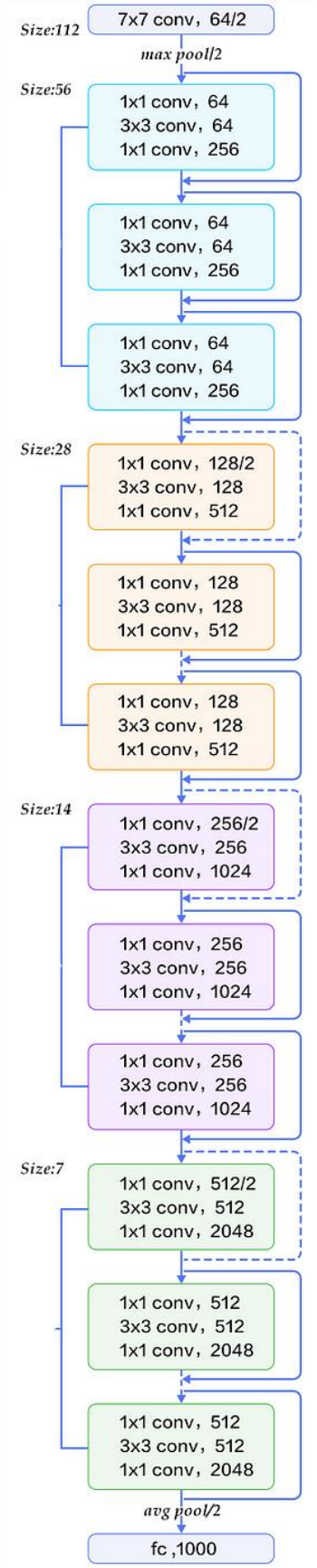
1. Stack the channels (t1, t2, t1ce, flair) on a new axis resulting in a 2D array.
2. Take the maximum value across that axis to combine the 2D array into a single channel image.
3. Using the maximum value per pixel per channel in the combination process to preserve the maximum intensity value from the available modalities.
4. Image is reshaped to match the expected input shape for the model.

This process ensures that for each patient slice, an image from one of the modalities is read and converted into a single-channel image, which is then used as input data for the model.

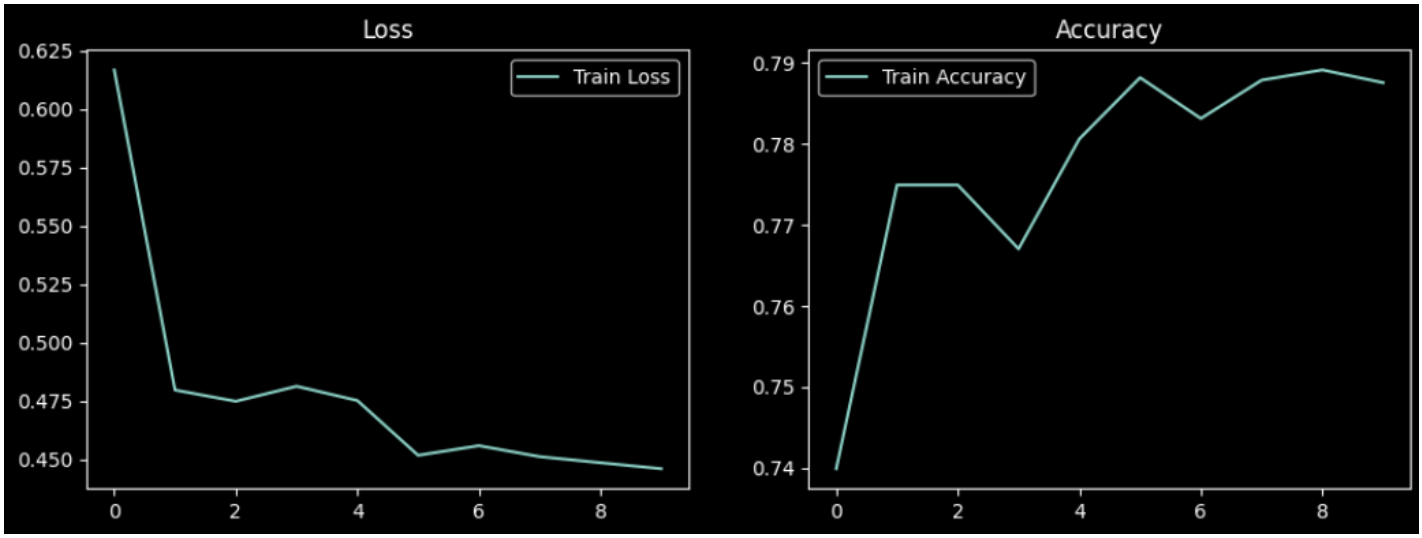
Data transformations

I. Resize
II. Combine channels

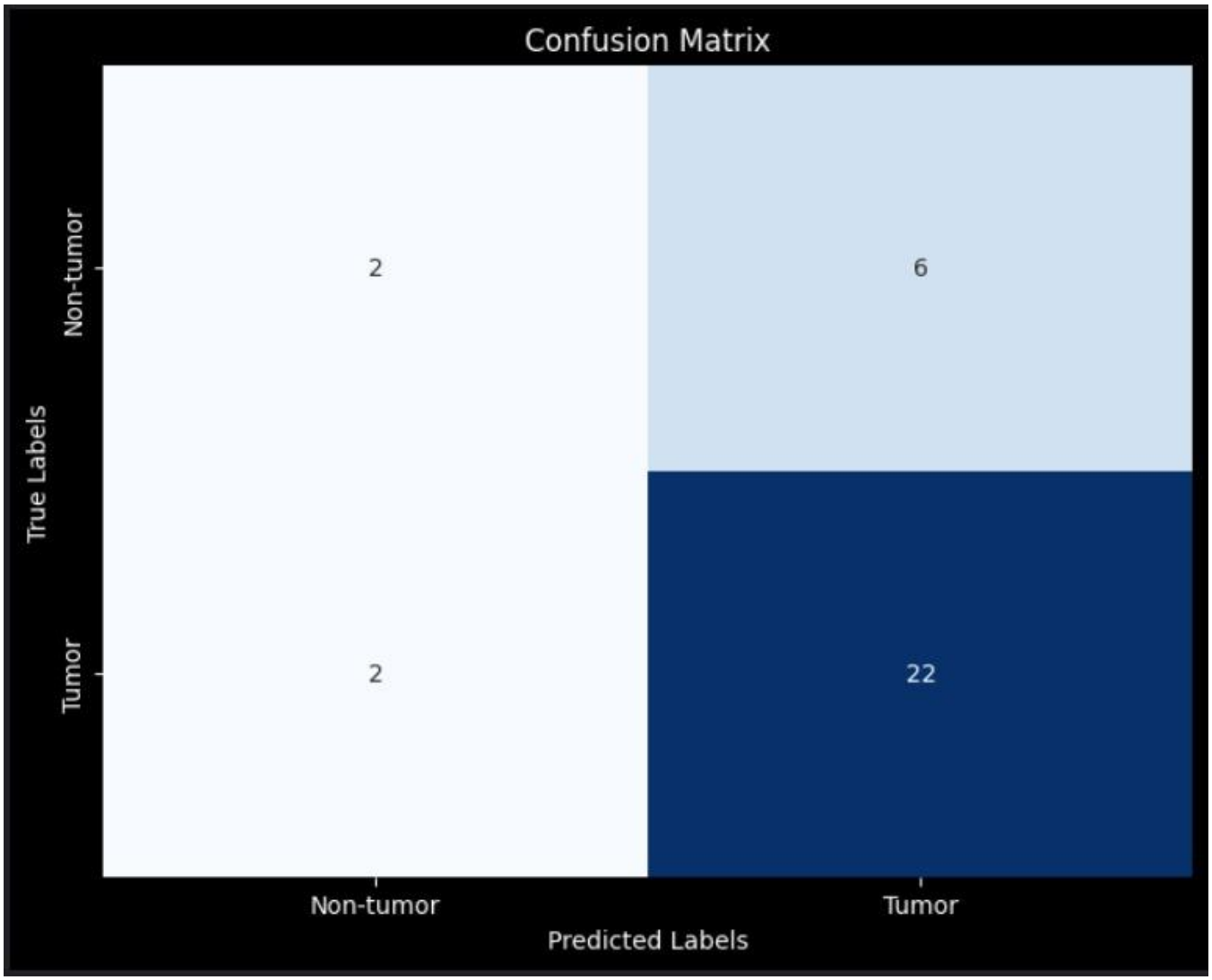
Model Architecture



Classification results



	precision	recall	f1-score	support
Non-tumor	0.50	0.25	0.33	8
Tumor	0.79	0.92	0.85	24
accuracy			0.75	32
macro avg	0.64	0.58	0.59	32
weighted avg	0.71	0.75	0.72	32


Confusion Matrix
Non-tumor
Tumor
True Labels
2
6
2
22
Non-tumor
Tumor
Predicted Labels

Interface

An interface is essential to visualize our model's capabilities to be used by end users not just programmers. We created a prototype for a web application that takes the MRI image from the user as input and output the segmentation and classification of the image. We used Flask for this task for multiple reasons outlined below.

Flask Advantages:

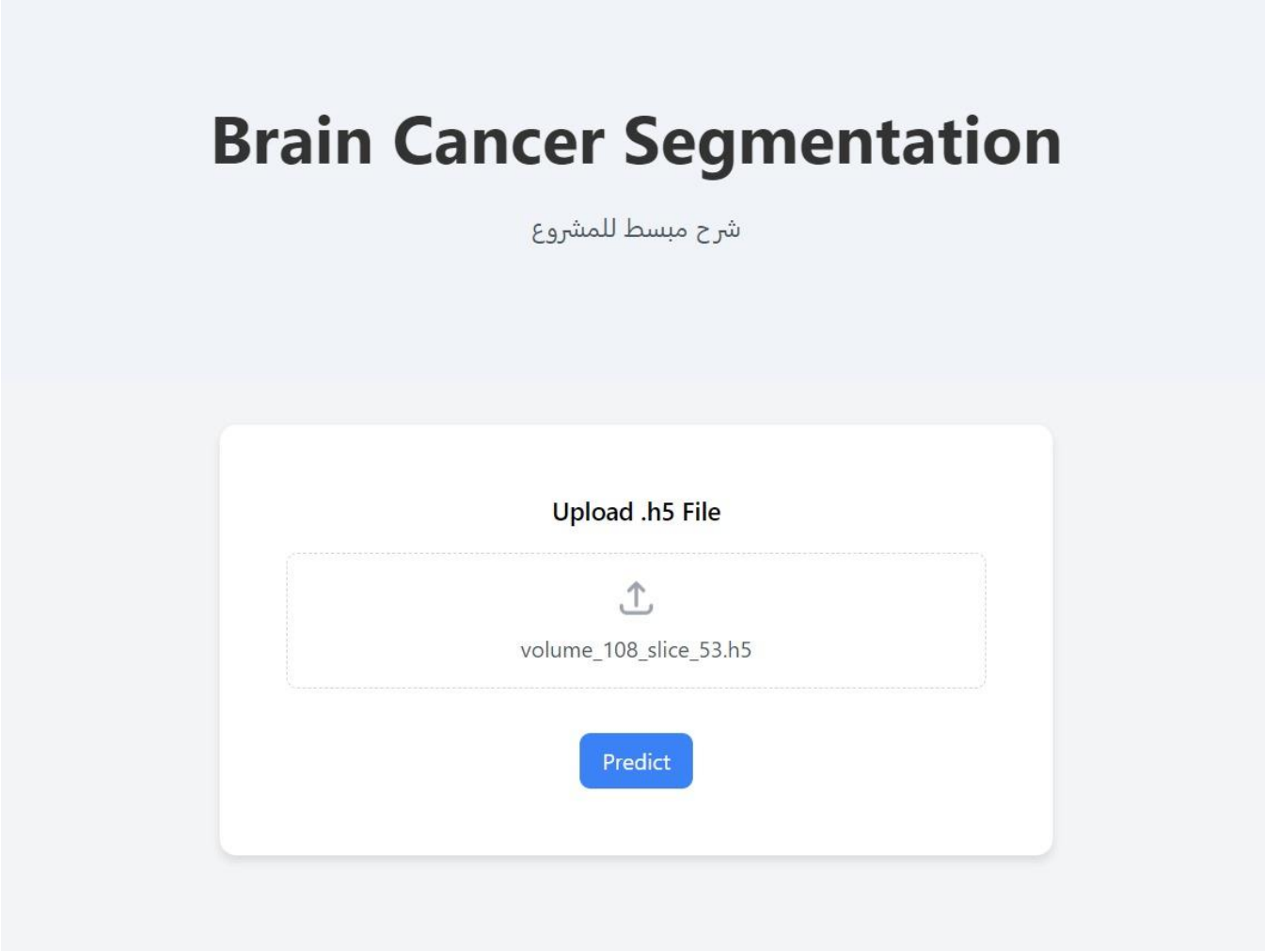
A. Simplicity and Ease of Use: Flask is lightweight and easy to set up, making it ideal for quickly developing and deploying classification models.

B. Flexibility: Flask is highly flexible and doesn't impose any constraints on how you structure your application. This flexibility is useful for customizing the backend to meet the specific needs of your classification project.

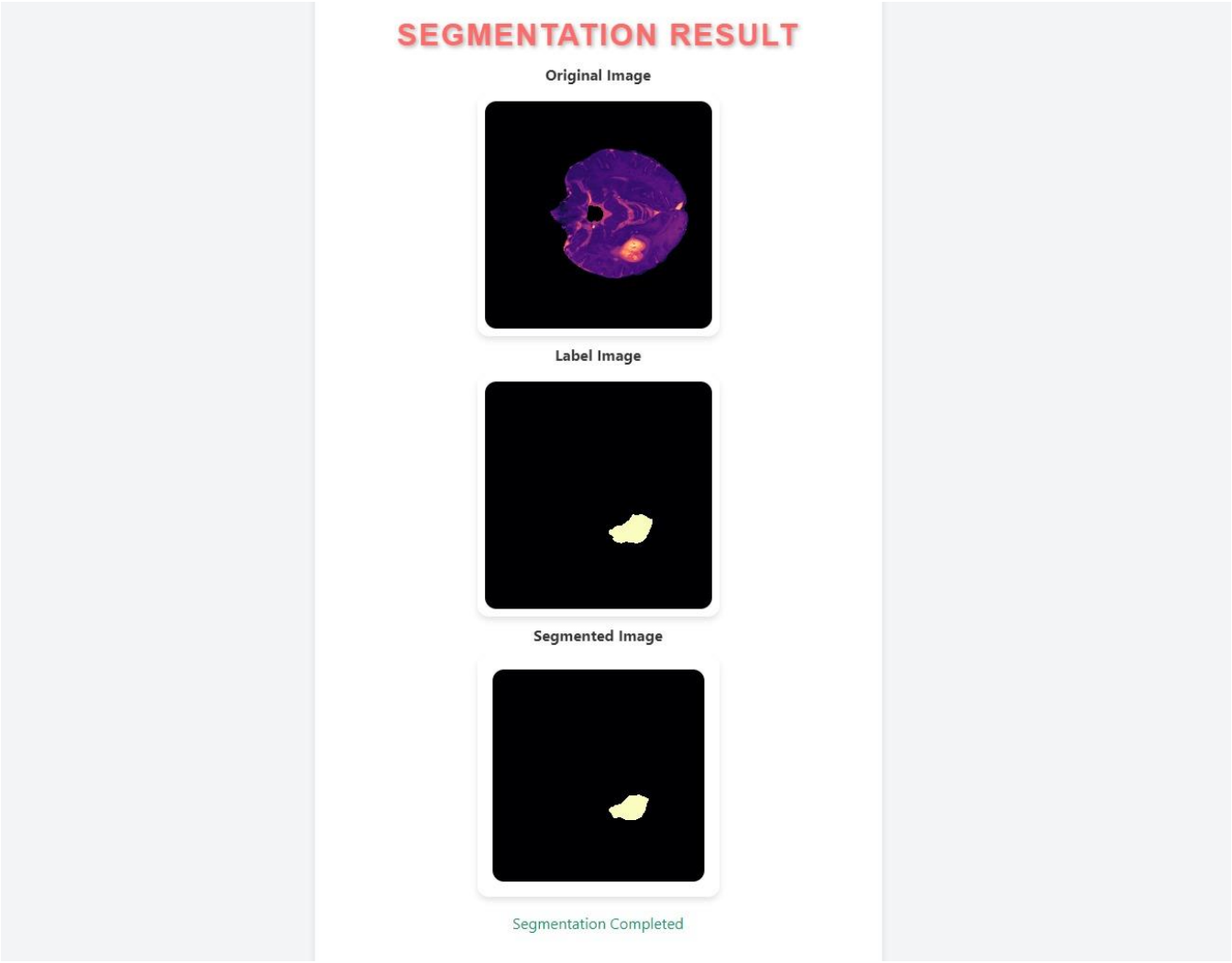
C. Integration with Machine Learning Libraries: Flask can be easily integrated with popular machine learning libraries like TensorFlow, PyTorch, and scikit-learn.

D. Rapid Development and Prototyping: Flask's minimalistic design allows for rapid development and prototyping, which is essential when testing and iterating on classification models.

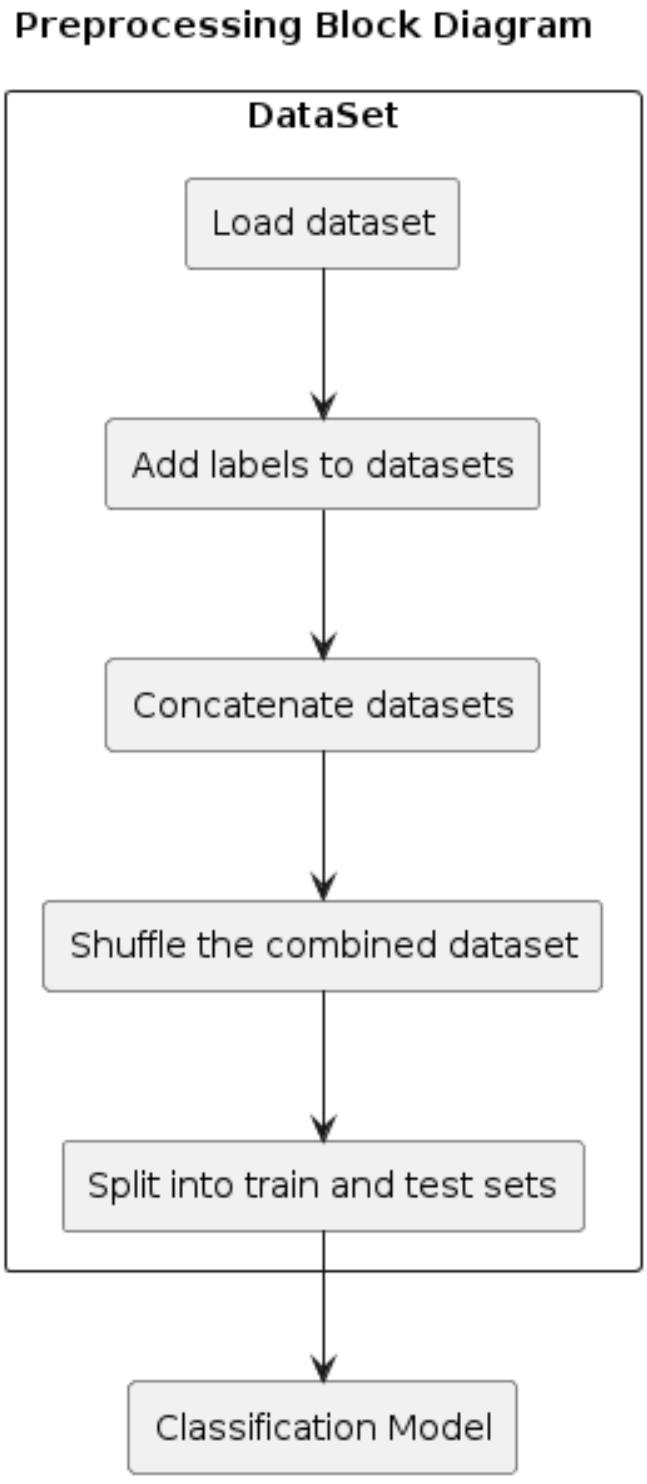
Main Screen

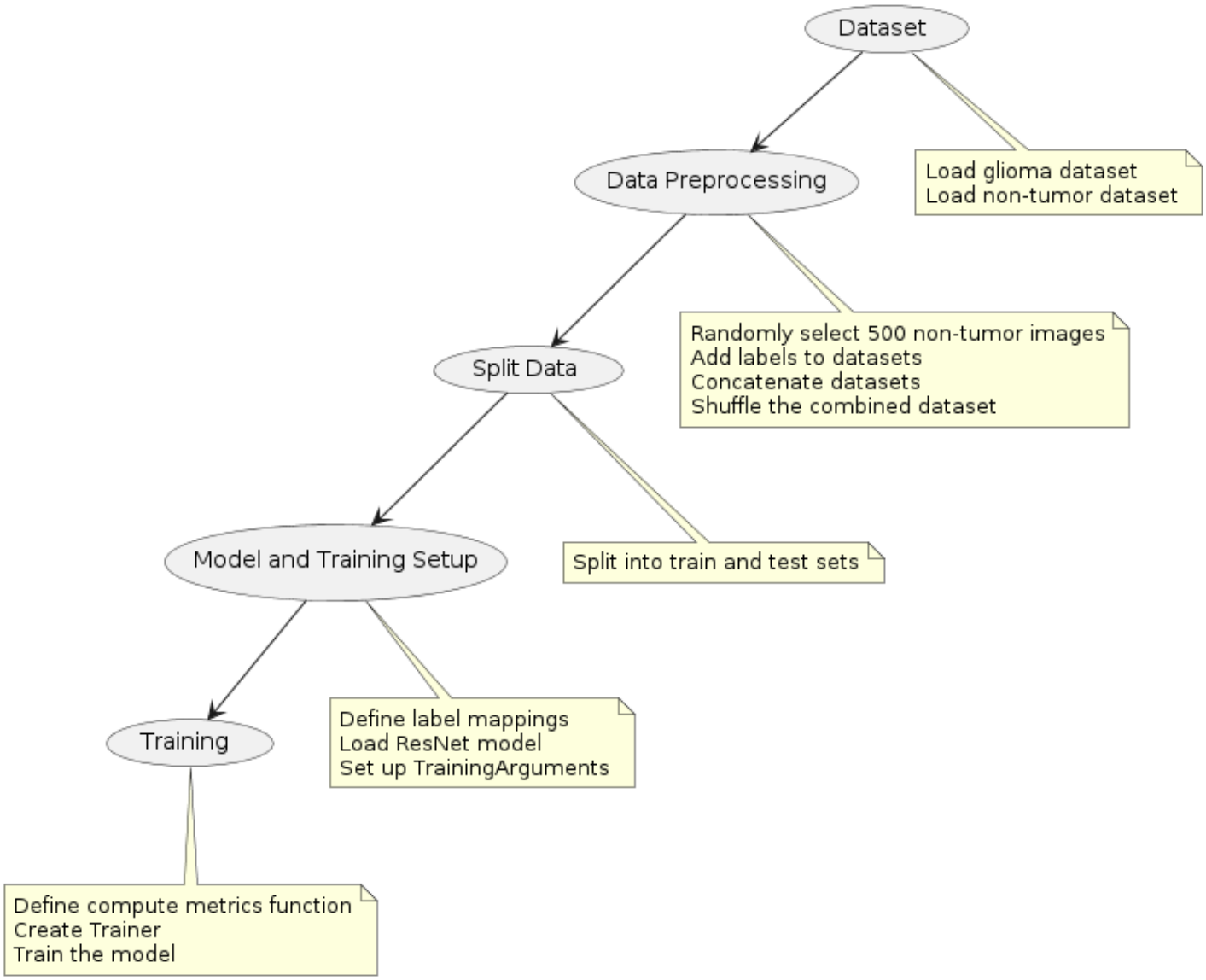

Brain Cancer Segmentation
شرح مبسط للمشروع
Upload .h5 File
volume_108_slice_53.h5
Predict

Example Output

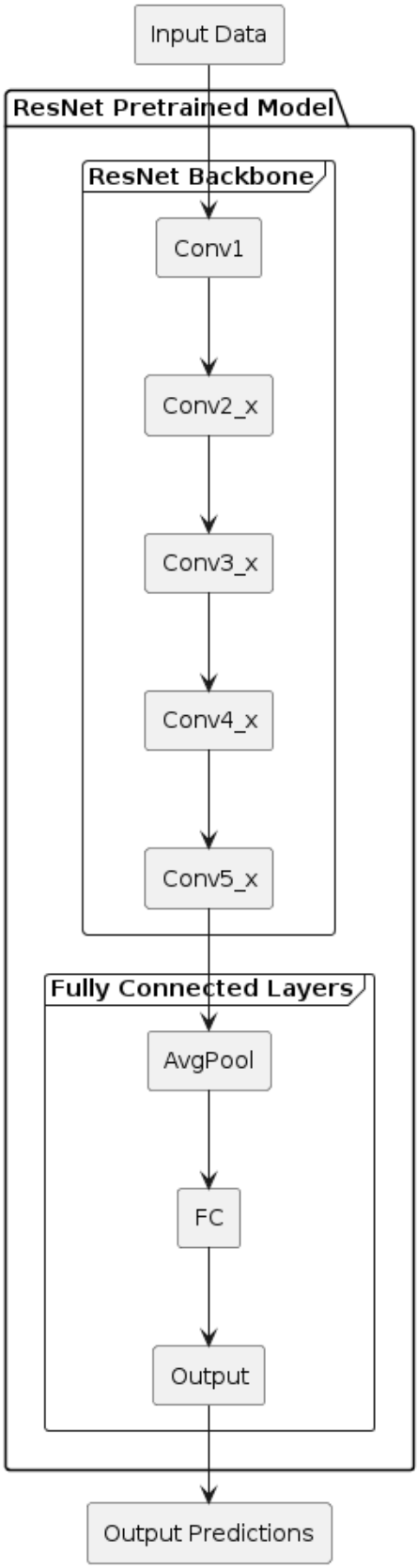


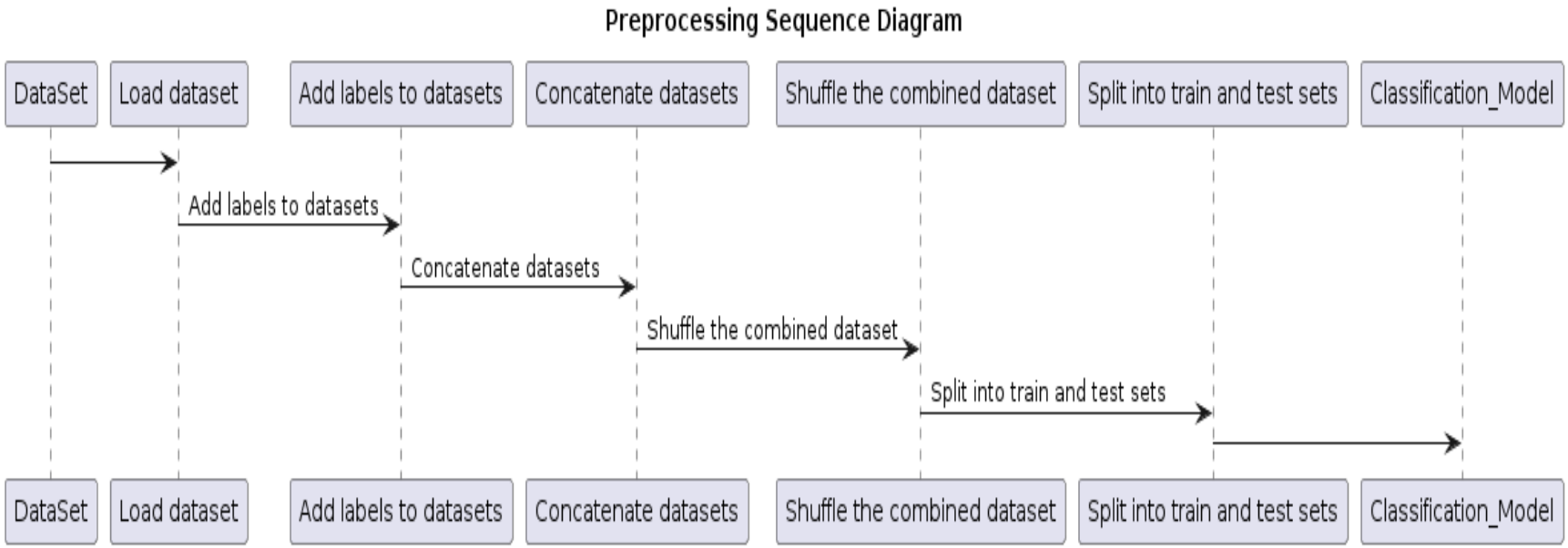
Diagrams

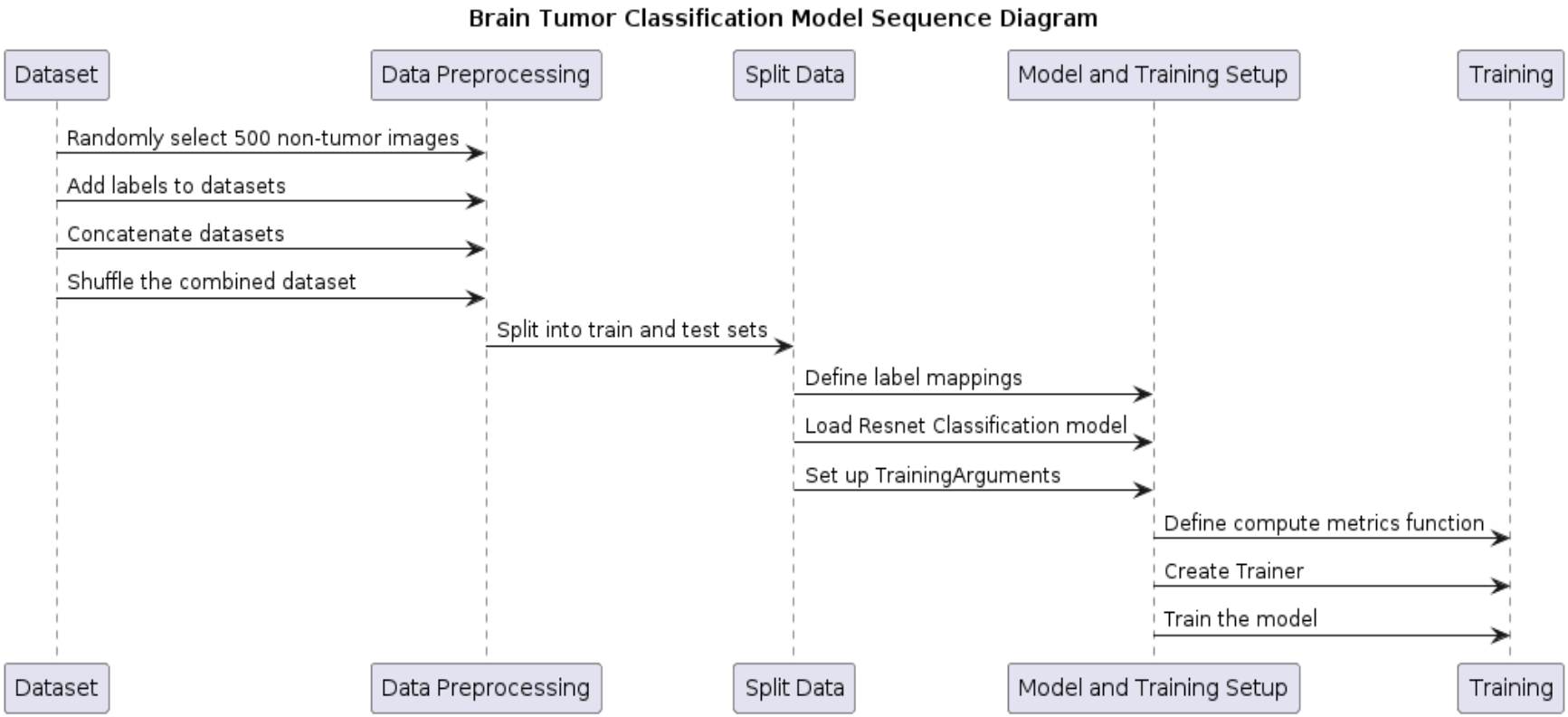

Preprocessing Block Diagram
DataSet
Load dataset
Add labels to datasets
Concatenate datasets
Shuffle the combined dataset
Split into train and test sets
Classification Model

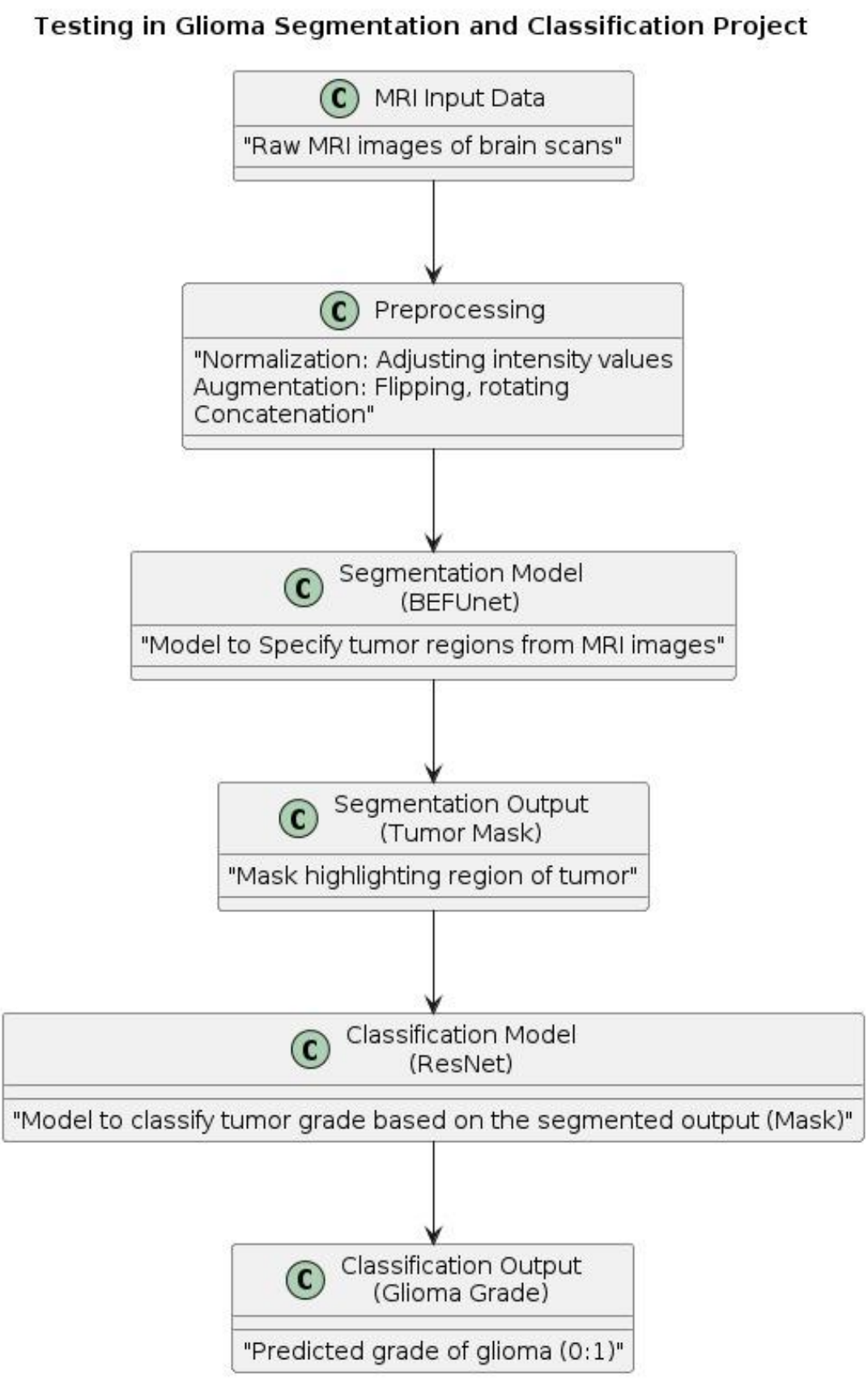

Brain Tumor Classification Model
Dataset
Data Preprocessing
Load glioma dataset
Load non-tumor dataset
Split Data
Randomly select 500 non-tumor images
Add labels to datasets
Concatenate datasets
Shuffle the combined dataset
Model and Training Setup
Split into train and test sets
Training
Define label mappings
Load ResNet model
Set up TrainingArguments
Define compute metrics function
Create Trainer
Train the model

Detailed ResNet Pretrained Model Block Diagram




Preprocessing Sequence Diagram
DataSet
Load dataset
Add labels to datasets
Concatenate datasets
Shuffle the combined dataset
Split into train and test sets
Classification_Model
Add labels to datasets
Concatenate datasets
Shuffle the combined dataset
Split into train and test sets
DataSet
Load dataset
Add labels to datasets
Concatenate datasets
Shuffle the combined dataset
Split into train and test sets
Classification_Model


Brain Tumor Classification Model Sequence Diagram
Dataset
Data Preprocessing
Split Data
Model and Training Setup
Training
Randomly select 500 non-tumor images
Add labels to datasets
Concatenate datasets
Shuffle the combined dataset
Split into train and test sets
Define label mappings
Load Resnet Classification model
Set up TrainingArguments
Define compute metrics function
Create Trainer
Train the model
Dataset
Data Preprocessing
Split Data
Model and Training Setup
Training


Testing in Glioma Segmentation and Classification Project
MRI Input Data
"Raw MRI images of brain scans"
Preprocessing
"Normalization: Adjusting intensity values
Augmentation: Flipping, rotating
Concatenation"
Segmentation Model
(BEFUnet)
"Model to Specify tumor regions from MRI images"
Segmentation Output
(Tumor Mask)
"Mask highlighting region of tumor"
Classification Model
(ResNet)
"Model to classify tumor grade based on the segmented output (Mask)"
Classification Output
(Glioma Grade)
"Predicted grade of glioma (0:1)"

Conclusion

Bef-Unet introduces a novel approach to semantic segmentation. The use of multi-headed attention allows the model to effectively and efficiently capture the important features of an image. We were able to with transfer learning and minimal training achieve an average mIoU of 0.805. This is a great result compared to the current state of the art approach, the EfficientNet FPN-based approach. This approach achieved an average mIoU of 0.887. For future work we intend to explore better approaches to further train the Bef-Unet model to reach better results.

References:

1) K. Han et al., "A Survey on Vision Transformer," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 1, pp. 87-110, 1 Jan. 2023, doi: 10.1109/TPAMI.2022.3152247.

2) Al-hammuri, K., Gebali, F., Kanan, A. *et al.* Vision transformer architecture and applications in digital health: a tutorial and survey. *Vis. Comput. Ind. Biomed. Art* **6**, 14 (2023)

3) Fahad Shamshad, Salman Khan, Syed Waqas Zamir, Muhammad Haris Khan, Munawar Hayat, Fahad Shahbaz Khan, Huazhu Fu,Transformers in medical imaging: A survey, Medical Image Analysis, Volume 88, 2023, 1361-8415

4) Yutong Bai and Jieru Mei and Alan Yuille and Cihang Xie, Are Transformers More Robust Than CNNs? , 2111.05464

5) Pagano, Tiago P. and Loureiro, Rafael B. and Lisboa, Fernanda V. N. and Peixoto, Rodrigo M. and Guimarães, Guilherme A. S. and Cruz, Gustavo O. R. and Araujo, Maira M. and Santos, Lucas L. and Cruz, Marco A. S. and Oliveira, Ewerton L. S. and Winkler, Ingrid and Nascimento, Erick G. S., Bias and Unfairness in Machine Learning Models: A Systematic Review on Datasets, Tools, Fairness Metrics, and Identification and Mitigation Methods, Big Data and Cognitive Computing, 2504-2289

6) Asiri, Abdullah A., Ahmad Shaf, Tariq Ali, Muhammad Ahmad Pasha, Muhammad Aamir, Muhammad Irfan, Saeed Alqahtani, Ahmad Joman Alghamdi, Ali H. Alghamdi, Abdullah Fahad A. Alshamrani, and et al. 2023. "Advancing Brain Tumor Classification through Fine-Tuned Vision Transformers: A Comparative Study of Pre-Trained Models" Sensors 23, no. 18: 7913. https://doi.org/10.3390/s23187913

7) Maciej A. Mazurowski, Haoyu Dong, Hanxue Gu, Jichen Yang, Nicholas Konz, Yixin Zhang, Segment anything model for medical image analysis: An experimental study,Medical Image Analysis, Volume 89, 1361-8415

8) Liu, Z., Tong, L., Chen, L. et al. Deep learning based brain tumor segmentation: a survey. Complex Intell. Syst. 9, 1001–1026 (2023). https://doi.org/10.1007/s40747-022-00815-5

9) Chen, Z., Peng, C., Guo, W. et al. Uncertainty-guided transformer for brain tumor segmentation. Med Biol Eng Comput (2023). https://doi.org/10.1007/s11517-023-02899-8

10) An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, Alexey Dosovitskiy and Lucas Beyer and Alexander Kolesnikov and Dirk Weissenborn and Xiaohua

Zhai and Thomas Unterthiner and Mostafa Dehghani and Matthias Minderer and Georg Heigold and Sylvain Gelly and Jakob Uszkoreit and Neil Houlsby, 201011929

11) Asiri, Abdullah A., Ahmad Shaf, Tariq Ali, Unza Shakeel, Muhammad Irfan, Khlood M. Mehdar, Hanan Talal Halawani, Ali H. Alghamdi, Abdullah Fahad A. Alshamrani, and Samar M. Alqhtani. 2023. "Exploring the Power of Deep Learning: Fine-Tuned Vision Transformer for Accurate and Efficient Brain Tumor Detection in MRI Scans" Diagnostics 13, no. 12: 2094.

12) Balayn, A., Lofi, C. & Houben, GJ. Managing bias and unfairness in data for decision support: a survey of machine learning and data engineering approaches to identify and mitigate bias and unfairness within data management and analytics systems. The VLDB Journal 30, 739–768 (2021)

13) Kshatri, S.S., Singh, D. Convolutional Neural Network in Medical Image Analysis: A Review. Arch Computat Methods Eng 30, 2793–2810 (2023). https://doi.org/10.1007/s11831-023-09898-w

14) Hardaha, S., Edla, D.R. & Parne, S.R. A Survey on Convolutional Neural Networks for MRI Analysis. Wireless Pers Commun 128, 1065–1085 (2023). https://doi.org/10.1007/s11277-022-09989-0

15) Spoerer, Courtney J. and McClure, Patrick and Kriegeskorte, Nikolaus, Recurrent Convolutional Neural Networks: A Better Model of Biological Object Recognition, Frontiers in Psychology, 1664-1078

16) A. B. Amjoud and M. Amrouch, "Object Detection Using Deep Learning, CNNs and Vision Transformers: A Review," in IEEE Access, vol. 11, pp. 35479-35516, 2023, doi: 10.1109/ACCESS.2023.3266093.

17) X. Li, L. Zhang, G. Cheng, K. Yang, Y. Tong, X. Zhu, and T. Xiang,"Global aggregation then local distribution for scene parsing," IEEE TIP, 2021.

18) Xiangtai Li, Henghui Ding, Haobo Yuan, Wenwei Zhang, Jiangmiao Pang, Guangliang Cheng, Kai Chen, Ziwei Liu, and Chen Change Loy, "Transformer-Based Visual Segmentation: A Survey", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, 2023.

19) Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," ECCV, 2020.

20) Yue Cao*, Jiarui Xu*, Stephen Lin, Fangyun Wei, Han Hu, "Global Context Networks", IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, 2020.

21) S. Das, O. F. M. R. R. Aranya and N. N. Labiba, "Brain Tumor Classification Using Convolutional Neural Network," 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), Dhaka, Bangladesh, 2019, pp. 1-5, doi: 10.1109/ICASERT.2019.8934603.

22) Hans Thisankea , Chamli Deshana , Kavindu Chamitha , Sachith Seneviratneb,c, Rajith Vidanaarachchib,c, Damayanthi Heratha,∗, Semantic segmentation using Vision Transformers: A survey, Engineering Applications of Artificial Intelligence, 2023.

23) Emerald U. Henrya*, Onyeka Emebob,c, Conrad Asotie Omonhinmind, Vision Transformers in Medical Imaging: A Review, 2022 , 2211.10043

24) Yuanduo Hong1 Jue Wang1 Weichao Sun1 Huihui Pan1∗,MINIMALIST AND HIGH-PERFORMANCE SEMANTIC SEGMENTATION WITH PLAIN VISION TRANSFORMERS, 2023, 2310.12755

25) Zhe Chen and Yuchen Duan and Wenhai Wang and Junjun He and Tong Lu and Jifeng Dai and Yu Qiao, Vision Transformer Adapter for Dense Predictions, 2023, 2205.08534

26) Qiang Wan and Zilong Huang and Bingyi Kang and Jiashi Feng and Li Zhang, Harnessing Diffusion Models for Visual Perception with Meta Prompts, 2023, 2312.14733

27) Yuanduo Hong and Jue Wang and Weichao Sun and Huihui Pan, MINIMALIST AND HIGH-PERFORMANCE SEMANTIC SEGMENTATION WITH PLAIN VISION TRANSFORMERS, 2023, 2310.12755

28) Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah. 2022. Transformers in Vision: A Survey. ACM Comput. Surv. 54, 10s, Article 200 (January 2022).

29) RSNA-ASNR-MICCAI Brain Tumor Segmentation (BraTS)

30) BEFUnet: A Hybrid CNN-Transformer Architecture for Precise Medical Image Segmentation, Omid Nejati Manzari and Javad Mirzapour Kaleybar and Hooman Saadat and Shahin Maleki, 2024.

31) A Comprehensive Guide to Microsoft's Swin Transformer