Segmentation of Polyps from Colonoscopy Image using Vision Transformer

03CS7914 Project (Phase II)

CHN20MT002 CHN20CSIP08 Vishnu Vinod
chn20csip08@ceconline.edu
M. Tech. Computer Science & Engineering (Image Processing)

Department of Computer Engineering
College of Engineering Chengannur
Alappuzha 689121
Phone: +91.479.2165706
http://www.ceconline.edu
hod.cse@ceconline.edu

College of Engineering Chengannur
Department of Computer Engineering

CERTIFICATE

This is to certify that, this report titled ***Segmentation of Polyps from Colonoscopy Image using Vision Transformer*** is a bonafide record of the work done by

CHN20MT002 CHN20CSIP08 Vishnu Vinod

Fourth Semester M. Tech. Computer Science & Engineering (Image Processing)

student, for the course work in **03CS7914 Project (Phase II)**, under our guidance and supervision, in partial fulfillment of the requirements for the award of the degree, M. Tech. Computer Science & Engineering (Image Processing) of **APJ Abdul Kalam Technological University**.

Guide

Princy Sugathan S
Asst. Professor
Computer Engineering

Coordinator

Ahammed Siraj K K
Associate Professor
Computer Engineering

Head of the Department

July 18, 2022

Dr. Manju S Nair
Associate Professor
Computer Engineering

Permission to Use

In presenting this project dissertation at College of Engineering Chengannur(CEC) in partial fulfillment of the requirements for a postgraduate degree from APJ Abdul Kalam Technological University, I agree that the libraries of CEC may make it freely available for inspection through any form of media. I further agree that permission for copying of this dissertation in any manner, in whole or in part, for scholarly purposes may be granted by the Head of the Department of Computer Engineering. It is understood that any copying or publication or use of this dissertation or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to CEC in any scholarly use which may be made of any material in this project dissertation.

Vishnu Vinod

Statement of Authenticity

I hereby declare that this submission is my own work and to the best of my knowledge it contains no materials previously published or written by another person, or substantial proportions of material which have been accepted for the award of any other degree or diploma at College of Engineering Chengannur(CEC) or any other educational institution, except where due acknowledgement is made in the report. Any contribution made to my work by others, with whom I have worked at CEC or elsewhere, is explicitly acknowledged in the report. I also declare that the intellectual content of this report is the product of my own work done as per the **Problem Statement** and **Proposed Solution** sections of the project dissertation report. I have explicitly stated the major references of my work. I have also listed all the documents referred, to the best of my knowledge.

Vishnu Vinod

Acknowledgements

The success and final outcome of any project requires a lot of guidance and assistance from many people and I am extremely fortunate to have all these along the completion of this project work. At the very offset of this project I am ineffable indebted to God Almighty for his showers of blessings on us for making this project a success.

With deep sense of reverence, I express my sincere thanks to **Dr. Smitha Dharan**, Principal, College of Engineering Chengannur, the Head of the Computer Science and Engineering Department **Dr. Manju S Nair**, Project Guide **Ms. Princy Sugathan S** and Project Coordinator **Mr. Ahammed Siraj K K** for extending all the facilities required for doing my project and conscientious guidance to accomplish this project.

I would like to express my gratitude to my parents for their constant moral and economic support. Last but not the least goes to all my friends who have directly or indirectly helped me for the completion of this project.

Vishnu Vinod

Abstract

Medical imaging is the technique of imaging the interior of a body for clinical analysis and medical intervention, as well as visual representation of the function of some organs or tissues. Image segmentation is an important medical image processing as it extracts the region of interest through a semiautomatic or automatic process. Accurate medical image segmentation is essential for the diagnosis and treatment planning of diseases. Convolutional Neural Networks (CNNs) have achieved state-of-the-art performance for automatic medical image segmentation and most of the research relies on them. So we have to try unfamiliar methods to expand the field of computer vision in medical image segmentation. Here a deep learning model called Vision Transformer is proposed for the segmentation of polyps from colonoscopy images.

Contents

List of Figures

Chapter 1

Introduction

Medical imaging, also known as radiology, is the field of medical science in which medical professionals recreate various images of parts of the body for diagnostic or treatment purposes. Different types of medical imaging techniques are X-rays, Magnetic resonance imaging (MRI), Ultrasounds, Endoscopy and Computerized tomography. Medical imaging allows us to assess patients' bones, organs, tissue and blood vessels through non-invasive means. Medical image segmentation is the processing of medical images that gives a better understanding of internal body architectures like their size, shape and orientation for treatment planning and surgeries, In addition, segmentation offers the benefit of removing any unwanted details from a scan, such as air, as well as allowing different tissues such as bone and soft tissues to be isolated.

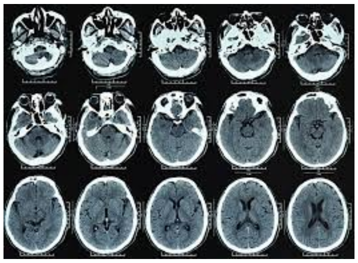

Figure 1.1: CT image

1.1 Polyp Segmentation from Colonoscopy Image

Colon cancer is one of dangerous cancer in the world that is found in our large intestine or colon. It begins as a small clump of cells called polyps. So the early detection of polyps is an important process to prevent the severity of colon cancer. Colonoscopy, a medical imaging technique is generally used for the detection of polyps, and via this technique images of affected colon areas can be captured and studied for diagnosis purposes. Segmentation of polyps from colonoscopy images is an important task in the process of polyp diagnosis and removal. By performing segmentation a better idea of characteristics of polyps like size, shape, and orientation can be acquired.

Figure 1.2: Colonoscopy image

Methods for performing segmentation vary widely depending on the specific application, imaging method, and other components. For example, the segmentation of lung cells from a chest CT scan image has different requirements from the segmentation of the polyps from a colonoscopy image. General imaging artifacts like noise, partial volume effects, and motion of scanning devices during image acquisition can also have significant effects on the performance of the segmentation method. Deep learning models like convolutional neural networks are mostly proposed and used for polyp segmentation. Convolutional neural networks outperform other traditional deep learning models and it is easy to understand and implement, therefore the major part of research and studies in this area is about convolutional neural networks. The application of other deep learning models in this field of polyp segmentation is inevitable to enhance the range of computer-aided support. So in this work, a deep learning model called vision transformer is proposed for the segmentation of polyps. Like convolutional neural networks, a vision transformer is a new approach to deep learning and is now widely used for medical image segmentation. The Vision Transformer, or ViT, employs a Transformer-like architecture over patches of the image. The core component of a vision transformer is the transformer encoder. It consists of a multi-head self-attention module, multi-layer perceptron module, and batch normalization layers.

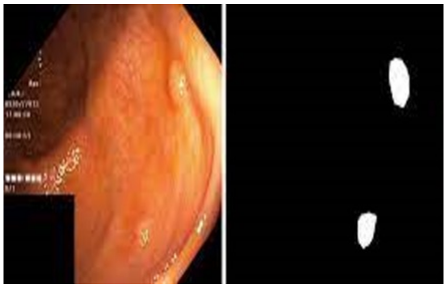

Figure 1.3: Colon Polyps Segmentation

1.2 Proposed Project

Aim of the project is to develop an automated framework which can segment the region of polyps from a colon image.

1.2.1 Problem Statement

Segmentation of colon polyps from colonoscopy image using a deep learning model.

1.2.2 Proposed Solution

Pyramid vision transformer with a convolutional neural network is proposed for the segmentation of colon polyps from colonoscopy image.

Chapter 2

Report of Preparatory Work

2.1 Literature Survey Report

Convolution-free medical image segmentation using transformers

In [1] they propose a convolution-free deep neural network for 3D medical image segmentation which uses only a transformer. The input to the model is a 3D block of the image. As an initial step, the block is partitioned into contiguous non-overlapping patches, then each of the patches is flattened into a vector of a specific size with positional embeddings. The encoder of this proposed network has a particular number of stages, each consisting of a multi-head self-attention layer and a subsequent two-layer fully connected feed-forward network. Both the MSA and FFN modules include residual connections, ReLU activations, and normalization. The proposed model is used to segment brain cortical plate from T2 MRI image, pancreas from CT image, and hippocampus from MRI image. They implemented the model in Tensorflow 1.16. In the first stage of training, they first pre-train the convolution free network 2 networks on unlabelled data for denoising and then fine-tune the network with labeled data of images. The dataset of brain cortical plate images contains 18 train images and 9 test images, the pancreas dataset contains 231 train images and 50 test images and the hippocampus dataset contains 220 train images and 40 test images.

A hybrid transformer architecture for medical image segmentation

A U-shaped hybrid transformer network is proposed for segmentation of various MRI scan images in [2]. The model integrates the strength of convolution and self-attention strategy for medical image segmentation. Here the idea behind convolution is to extract local intensity features and the transformer self-attention is used to capture long-range associative information from the input MRI image. The proposed transformer network block in this study uses multi-head self-attention of four heads and the output from that attention module is the scaled dot product of key, query, and value vectors organized with positional embeddings corresponding to the input MRI images. They apply the transformer block to each level of the encoder and decoder modules. The encoder and decoder are a set of connected residual blocks that are composed of two convolutional layers with ReLU activation function and two batch normalization layers. In the training dataset there are 150 MRI annotated images and in the tesing dataset there are 200 annotated images.

TransBTS: Multimodal brain tumor segmentation using transformer

In [3] they present a transformer with a convolutional neural network for 3D MRI brain tumor segmentation. The proposed method is based on an encoder-decoder structure. The encoder is a 3D convolutional neural network with four convolution layers which is used to generate compact feature maps from input MRI images with spatial and depth information. The decoder is also a 3D convolutional neural network composed of four deconvolution layers and it performs feature upsampling and pixel-level segmentation. Convolution and deconvolution layers for downsampling and features within the encoder and upsampling the features within the decoder respectively are connected via skip connections. In between the output layer of the encoder and the input layer of the decoder, they place the transformer encoder. The transformer encoder is composed of a certain number of transformer layers which consists of a multi head self attention block and feed forward network block. The feature maps from the encoder block will be converted into a set of patches with posiyional embeddings and then given as input tho the transformer encoder. Here the proposed method processes each 3D medical image in a slice by slice manner. They use the Brain Tumor Segmentation 2019 challenge (BraTs) dataset contains 335 cases of patients for training and 125 cases of patients for testing the network. Each sample in the dataset is composed of T1-weighted modality, post-contrast T1-weighted modality, T2-weighted modality, and fluid-attenuated inversion recovery modality of MRI brain scans. Each modality image has a volume of 240*240*155.

U-Net Transformer: self and cross attention for medical image segmentation

In [4] an encoder-decoder-based U-shaped transformer model with two types of attention modules, multi-head self-attention, and multi-head cross attention is proposed. The encoder of this model is composed of four convolution layers and in between each pair of convolution layers, a maxpooloig layer is connected. The overall function of the encoder is the downsampling of input image features. The decoder is composed of three convolution layers with maxpooling layers for upsampling the features. The multi-head self-attention module is placed at the end of an encoder and it is designed to extract long-range structural information from the images. The multi-head cross attention module is placed within the skip connections between encoder and decoder layers to turn off irrelevant or noisy areas from the skip connection features. The proposed model is used for segmentation of abdominal organ using TCIA pancreas public dataset for training and validation of the model.

A transformer-based network for anisotropic 3D medical image Segmentation

In [5] a transformer-based network is proposed for segmentation of lungs from 3D chest CT scan image. Here the model uses 2D Unet as a back borne. It consists of a down-sampling encoder, an up-sampling decoder, and a transformer block in between the encoder and decoder. The feature maps from the down-sampling layers are forwarded to the up-sampling layers via skip connections. At the end layer of the encoder, the feature maps are sampled and passed to the transformer. Here in the transformer module, there are three convolutional layers that are used to make queries, keys, and values from the input sequence of features then positional embeddings are added to them. The network is trained using the lung cancer segmentation dataset published by The Medical Segmentation Decathlon. It contains 20,707 CT scan images of lungs.

UNETR: Transformers for 3D medical image segmentation

In [6] the proposed model is made up of a transformer encoder and a convolutional neural network-based decoder. The encoder and decoder are connected via skip connections. The model is used for 3D CT images and MRI image segmentation. As an initial step, the input images are divided into a sequence of patches with positional embeddings. And passed to the transformer encoder block. The transformer encoder is composed of multi-head self-attention and multilayer perceptron sublayers. A multi-head self-attention layer comprises a specific number of attention heads. The features from multiple resolutions of the encoder are merged with the decoder network to extract a region of interest from images. The decoder is made up of 6 convolution layers and 3 deconvolution layers and each of the convolution and deconvolution layers is connected with a batch normalization layer and a ReLu activation layer. Here BTCV dataset consists of an abdominal CT scan image of 30 patients and the MSD dataset consists of a brain MRI image for brain tumor segmentation of 484 patients are used for training the of the proposed model.

HT-net: hierarchical context-attention transformer network for medical ct image segmentation

In [7] a hierarchical context-attention transformer network with convolutional neural network as backbone is proposed for 2D CT image segmentation. Here pre-trained ResNet34 is used as the encoder-decoder network. In all skip connections between convolution layers of encoder and decoder modules, the transformer block is connected. The transformer is composed of three modules , a residual atrous pyramid pooling (RAPP)module, hierarchical context attention (HCA), and a position-sensitive axial attention (PAA) module. The RAPP module sets different combinations of dilation rates according to the size of feature maps at different encoder stages, adapts the size of the feature map at the current stage, avoids overlapping of the functions of some branches, and obtains more diverse receptive fields. Secondly, for leveraging the transformer mechanism to obtain self-attention for multiscale samples, in each stage of skip-connection, a position-sensitive axial attention (PAA) module is used following RAPP module. Thirdly, a hierarchical context attention (HCA) module to compensate for the global context-attention information lost by the long-range associations between image patches modeled by the transformer mechanism. In each skip-connection, the RAPP module, the PAA module, and the HCA module are added sequentially. The proposed network is used for segmentation of bladder segmentation, kidney segmentation, and lung segmentation from various CT images. For bladder segmentation, the model is trained on a dataset of CT images with 3520 images for training and 880 images for testing. KiTS19 dataset is used for training the model for kidney segmentation, which contains 6950 images with ground truths. The model is trained on a lung dataset comes from the Lung Nodule Analysis (LUNA) competition and Data Science Bowl 2017 and contains 534 2D CT images with respective label images.

Swin-unet: Unet-like pure transformer for medical image segmentation

In [8] also a convolution free neural network is proposed for 2D medical image segmentation, which is made up of swin transformer blocks. The model has a U-shaped architecture consisting of encoder, decoder, and skip connections. Encoder and decoder are Swin transformer blocks. Swin transformer is constructed with a window-based multi-head attention layer, normalization layer, and multi-layer Perceptron network. In the first step, the medical images are split into non-overlapping patches and positional embeddings are added to the patches and passed to the encoder for extracting features. With the encoder, there is a patch merging layer for downsampling the feature and increasing the dimension. In the decoder, the extracted context features are fused with multiscale features caused by downsampling. And a patch expanding layer with the decoder will reshape the feature maps into larger feature maps by upsampling the resolution. The proposed model is trained with a synapse multi-organ segmentation dataset which contains 30 cases with 3779 axial abdominal clinical CT images.

2.2 System Study Report

The output of the proposed method for segmentation is a new image which only contains the regions that we need from the input medical image. In the case of polyp segmentation the resulting image will only contain the region of polyps. The proposed model must be trained with a dataset of medical images and corresponding segmentation masks(Pre-defined segmentation output).

Chapter 3

Project Design

3.1 Resource Requirements

3.1.1 Hardware & Software Requirements

Processor : Intel Core i5 7th Gen 3.2GHz
Supporting Software and Libraries : Python,
Tensorflow, Keras, OpenCv , PyTorch
RAM : 12GB
Graphics Card : 6GB NVIDIA GeForce GTX 1060 GPU
Operating System : Any Operating System
Supporting Environment : Google Colab
and PC available in Computer Lab-4 CEC

3.1.2 Data Requirements

Here two datsets are needed, one for training the model and the other for testing the model.
For training and testing, the Kvasir-SEG dataset is chosen. The Kvasir-SEG dataset (size 46.2 MB) contains 1000 polyp images and their corresponding ground truth from the Kvasir Dataset v2. The resolution of the images contained in Kvasir-SEG varies from 332x487 to 1920x1072 pixels. The images and its corresponding masks are stored in two separate folders with the same filename. The image files are encoded using JPEG compression, and online browsing is facilitated. 900 images from the dataset are used for training and 100 images are used for testing.

3.2 Method

Most of the deep learning approaches for medical image segmentation have an encoder-decoder architecture. The goal of the encoder module is the extraction of salient features from the input medical image with the help of various layers like the convolution layer, max-pooling layer, and activation layer. And the decoder module is also composed of convolution layers and other neural layers for concatenation of encoded features from the encoder to produce the segmented image. Here a vision transformer model called Pyramid vision transformer is used as the encoder and a simple convolutional neural network is used as the decoder. A color image (RGB image with 3-channels) will be given as input to the encoder of the proposed model. Features are extracted with

in the four levels of the pyramid vision transformer. The encoder can produce 64, 128, 350 and 512 feature maps of the input image using the first, second, third and fourth level of pyramid vision transformer respectively. After the feature extraction by encoder, the features extracted will be combined and decoded using the decoder. An input image to a vision transformer will go through the following steps:
1. Split an image into patches
2. Flatten the patches
3. Produce lower-dimensional linear embeddings from the flattened patches
4. Add positional embeddings
5. Feed the sequence as an input to a standard transformer encoder

.

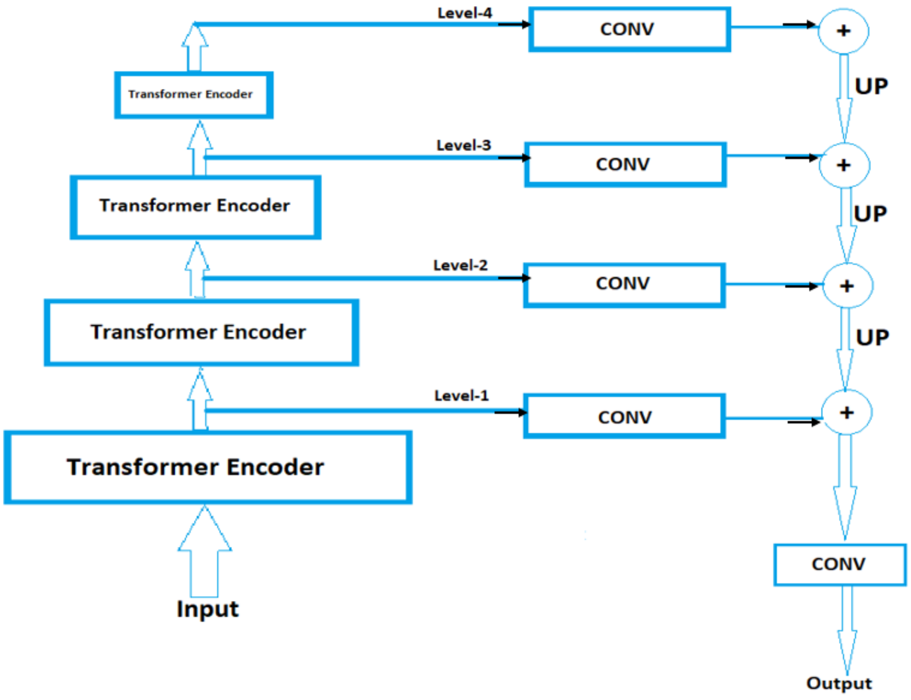

Figure 3.1: Model

3.2.1 Encoder: Pyramid Vision Transformer

The entire model is divided into four stages, each of which is comprised of a patch embedding layer and a Li-layer Transformer encoder. The patch embedding layer will divide the input image into a specific number of patches or blocks with same size. And these patches with positional embeddings will passed to first level transformer encoder of the model. The core components of transformer encoder are multihead self attention layer, normalization layer and multi layer Perceptron.

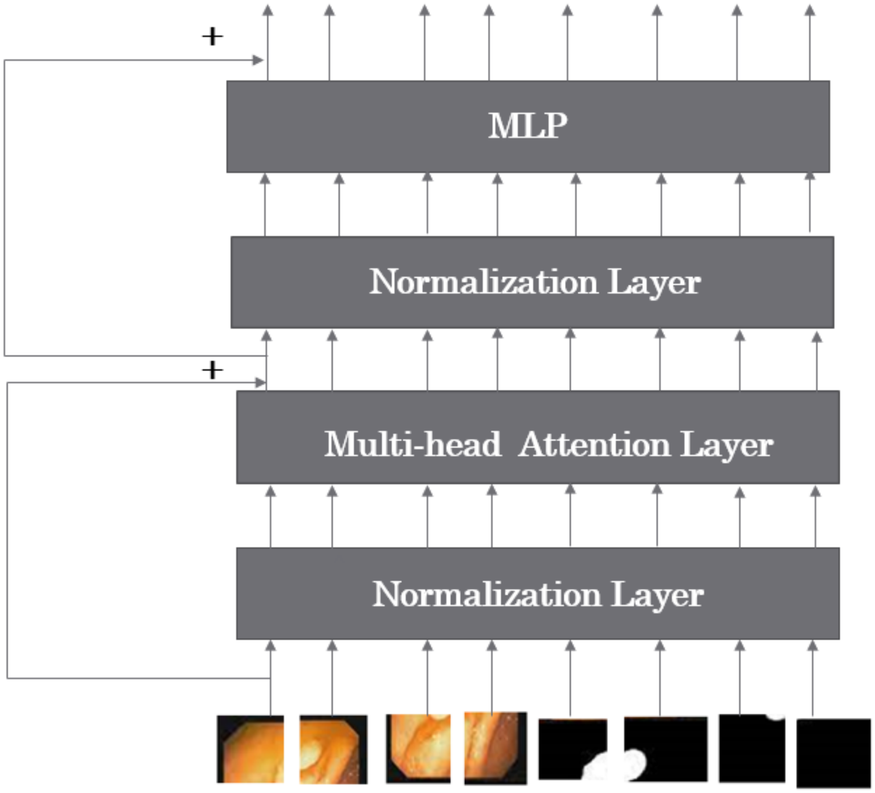

Figure 3.2: Transformer Encoder

Multi Head Self Attention Layer

Multi-head Attention is a module for attention mechanisms that run through an attention mechanism several times in parallel. The independent attention outputs are then concatenated and linearly transformed into the expected dimension. As an initial step, the input image patches with positional embeddings will be partitioned into queries, values and keys. To obtain these representations, every input is multiplied by a set of weights for Keys (denoted as K), a set of weights for queries (denoted as Q), and a set of weights for values (denoted as V).

General working of multihead attention is as follows.

1. Compute the linearly projected versions of the queries, keys and values through a multiplication with the respective weight matrices for each head.

2. Apply the single attention function for each head by multiplying the queries and keys matrices, applying the scaling and softmax operations(weighting the values matrix, to generate an output for each head).

3. Concatenate the outputs of the heads.

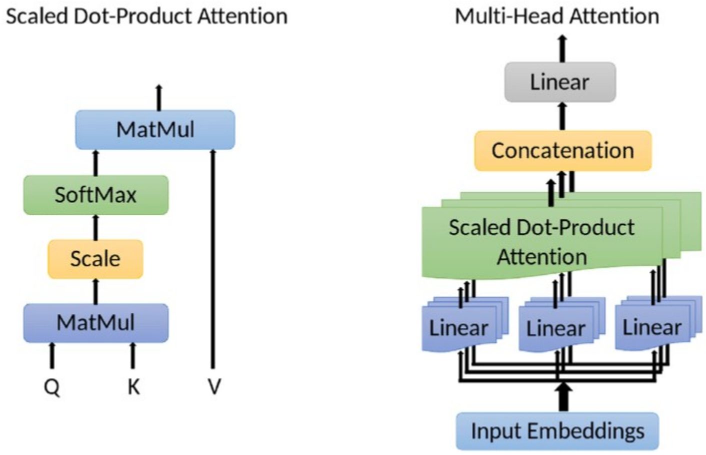

Figure 3.3: Attention Layer

Normalization Layer

Normalization is a method usually used for preparing data before training the model. The main purpose of normalization is to provide a uniform scale for numerical values. If the dataset contains numerical data varying in a huge range, it will skew the learning process, resulting in a bad model. The normalization method ensures there is no loss of information and even the range of values isn't affected.

$$y = \frac{x - \mathrm{E}[x]}{\sqrt{\mathrm{Var}[x] + \epsilon}} * \gamma + \beta$$

The mean and standard-deviation are calculated for data inputs. and are learnable affine transform parameters set to 0 and 1.

Multi Layer Perceptron

Multilayer perceptron (MLP) is a feed-forward neural network model. MLP contains a dropout layer, linear layer, an activation layer and a convolution layer. Dropout technology will randomly stop a certain number of neurons in the hidden layer, and use the mask process to set the output of these neurons in the hidden layer to 0, while the connection weights of the non-working neurons will not be updated in this iteration process. And the linear layer use matrix multiplication to transform their input features into output features using a weight matrix. The input features are received by a linear layer are passed in the form of a flattened one-dimension tensor and then multiplied by the weight matrix. And there is an activation layer with GELU activation function in MLP.

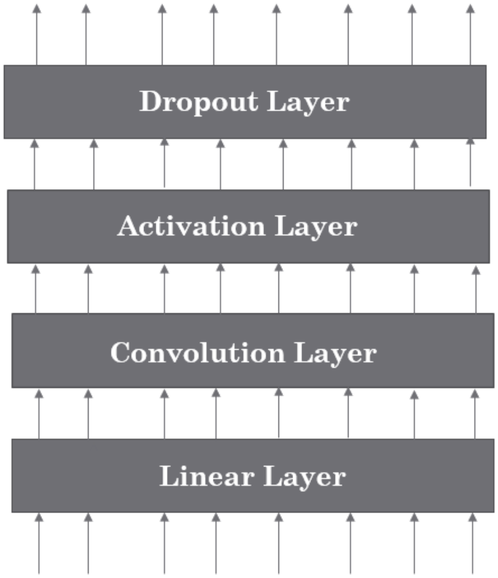

Figure 3.4: mlp

Spatial Reduction Attention

When compared to normal transformer encoder the encoder of a pyramid vision transformer has a spatial-reduction attention (SRA) layer with the multihead self attention layer. SRA uses average pooling to reduce the spatial dimension (h × w) to a fixed size (P × P) before the attention operation. Average Pooling calculates the average value for patches of a feature map, and uses it to create a downsampled (pooled) feature map.

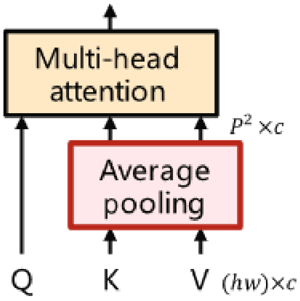

Figure 3.5: SRA

3.2.2 Decoder

The basic operation performed by this module is the addition of four levels of feature maps 64, 128, 320, 512 from the encoder with a size of 88×88, 44×44, 22×22 and 11×11 respectively. For that, as an initial stage, the four-level features are converted to 32 feature maps using convolution units, and then higher-level features are upsampled to the size of lower-level features. The feature maps in level-4 are upsampled to the size of level-3 feature maps then both are added, after that the resultant features of level-4 and level-3 addition are upsampled and then added with feature maps of level-2. Then the result of level-2 addition will be added with the feature maps of level-1 after upsampling. After adding the feature maps together the result will be passed through a single output convolution layer to produce the final output image of segmentation by reducing the dimension.

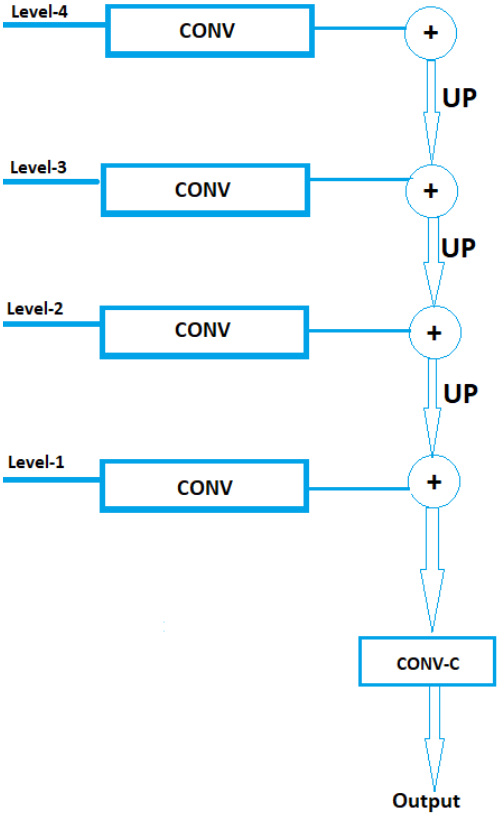

Figure 3.6: Decoder

Consider L1, L2, L3, and L4 as the features from four levels of the pyramid vision transformer encoder. After adjusting the number of feature maps to 32 via convolution units they become l1, l2, l3 and l4. l4 is upsampled to the same size of l3 and then added with it as l34=l3+up(l4). Then l34 is upsampled to the size of l2 and then added with it as l23=l2+up(l34). And then l23 is upsampled to the same size of l1 and then added with it as l12=l1+up(l23). In the final stage of decoder the resultant set of feature maps will be fed to a simple convolution layer C to reduce the dimension to single feature map.

3.3 Dataset and Training

Training of the proposed deep learning model is an important task. In the case of a segmentation problem the proposed model must be trained with a dataset of images and their corresponding segmentation masks. Here Kvaser-SEG dataset is used.

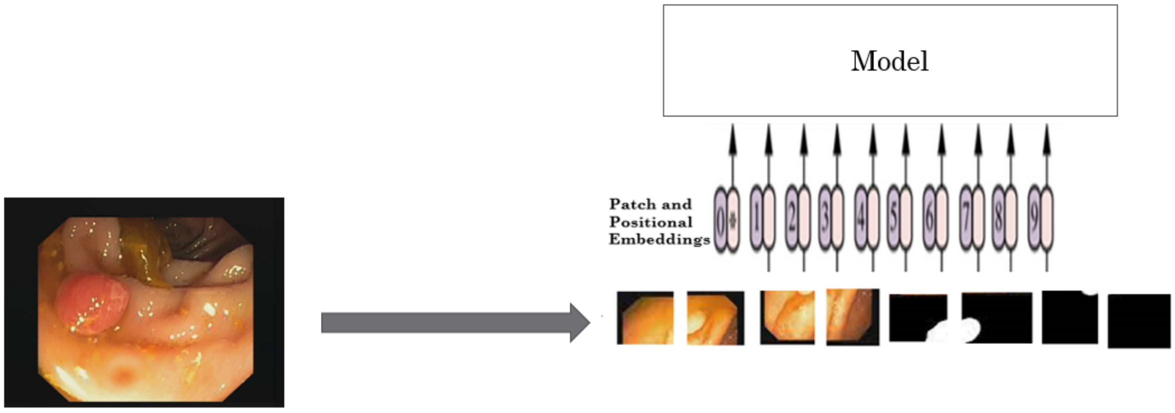

Figure 3.7: Training of ViT

3.4 Segmentation using Trained Model

The outcome of training process will be a vision transformer model for colonoscopy image segmentation. And when we input a colonoscopy image to the model the output will be the segmented polyps. The input image to the vision transformer should be in a specific size, for colonoscopy image the size is 352 × 352. So the input should be resized as a pre-processing step.

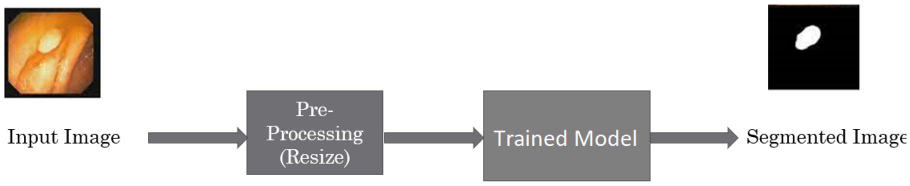

Figure 3.8: Segmentation

Chapter 4

Implementation

The proposed model is implemented in Pytorch framework using Google Colab platform. The model is then trained with Kavaser-Seg dataset[26] with a batch size of 16 and image size of 352 × 352. Here AdamW optimizer is used as the optimizer for training and the learning rate is 1e-4. Here the loss function used is IoU loss. After training, the model is evaluated with some of polyp images with known segmentation output.

Optimizer	Learning Rate	Batch Size
AdamW	1e-4	16
Epochs	Image Size	Loss Function
100	352 × 352	IoU Loss

Table 4.1: Training Parameters

Chapter 5

Results & Conclusions

Medical image segmentation is an important computer vision task in medical field. For treatment planning of diseases and computer aided surgeries the region of affected organ must be segmented from the medical image. Here a deep learning model which consist of pyramid vision transformer as encoder and a convolutional neural network as decoder is proposed.

5.1 Performance Analysis of Model

Four widely-used evaluation metrics, including Dice, IoU, precision and recall is used to evaluate the model performance by testing the model using Kavaser Seg dataset. Among these metrics, Dice and IoU are similarity measures at the regional level, which mainly focus on the internal consistency of segmented objects.

Mean Dice	Mean IOU	Precision	Recall
0.78461	0.77092	0.9112	0.73462

Table 5.1: Evaluation Results

Some of the results of testing of the model is shown below.

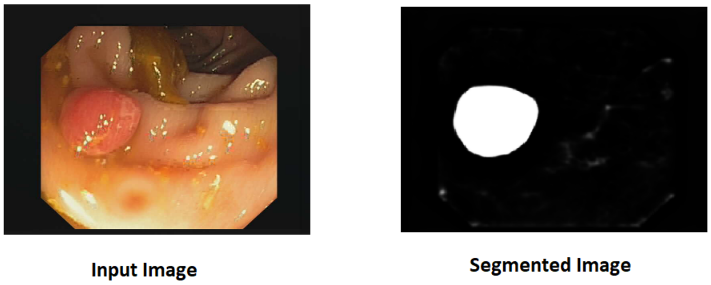

Figure 5.1: Input and Output

5.2 Conclusion

I studied a lot of research papers and other resources based on medical image segmentation and it is found that, most of the researches relies on convolutional neural networks. I proposed a method called vision transformer for segmentation of polyps, which is the current trend in medical applications. When compared to other deep learning models like convolutional neural network the proposed model is going to use the actual input image without any preprocessing.

5.3 Publication

A review paper named 'Medical Image Segmentation Using Vision Transformers' is presented in IEEE World Conference on Applied Intelligence and Computing (AIC 2022) Organized by Rajkiya Engineering College Sonbhadra(India), Technically supported by Soft Computing Research Society, June 17-19, 2022.

5.4 Future Scope

- In future different image preprocessing techniques can be applied in the input image before actual segmentation by the model to analyze the effect of preprocessing in segmentation.
- Here the model is trained in normal system without GPU, so the performance of the model can be improved by training the model in a high performance system.
- With more studies and researches the proposed segmentation model can be used for the segmentation of other medical images like CT image, MRI image, etc.

References

[1] Convolution-Free Medical Image Segmentation Using Transformers, Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, 2021, Volume 12901,ISBN : 978-3-030-87192-5, Davood Karimi, Serge Didenko Vasylechko, Ali Gholipour

[2] UTNet: A Hybrid Transformer Architecture for Medical Image Segmentation, Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, 2021, Volume 12903, ISBN : 978-3-030-87198-7, Yunhe Gao, Mu Zhou, Dimitris N. Metaxas

[3] TransBTS: Multimodal Brain Tumor Segmentation Using Transformer, Medical Image Computing and Computer Assisted Intervention – MICCAI 2021, 2021, Volume 12901, ISBN : 978-3-030-87192-5, Wenxuan Wang, Chen Chen, Meng Ding

[4] U-Net Transformer: Self and Cross Attention for Medical Image Segmentation, Machine Learning in Medical Imaging, 2021, Volume 12966 ISBN : 978-3-030-87588-6, Olivier Petit, Nicolas Thome, Clement Rambour

[5] D. Guo and D. Terzopoulos, ”A Transformer-Based Network for Anisotropic 3D Medical Image Segmentation,” 2020 25th International Conference on Pattern Recognition (ICPR), 2021, pp. 8857-8861, doi: 10.1109/ICPR48806.2021.9411990.

[6] A. Hatamizadeh et al., ”UNETR: Transformers for 3D Medical Image Segmentation,” 2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 1748-1758, doi: 10.1109/WACV51458.2022.00181.

[7] Xia H. Tan Y Ma, M. Ht-net: hierarchical context-attention transformer network for medical ct image segmentation. Appl Intell (2022).

[8] u Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xi- aopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmenta- tion, 2021.

[9] Bo Dong, Wenhai Wang, Deng-Ping Fan, Jinpeng Li, Huazhu Fu, and Ling Shao. Polyp-pvt: Polyp seg- mentation with pyramid vision transformers. CoRR, abs/2108.06932, 2021.

[10] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: Endto-end object detection with transformers. In: European Conference on Computer Vision. pp. 213–229. Springer (2020)

[11] C¸ i¸cek, O., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3d u-net: ¨ Learning dense volumetric segmentation from sparse annotation. In: MICCAI. pp. 424–432 (2016)

[12] Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.04805 (2018), http://arxiv.org/abs/1810.04805

[13] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2021)

[14] Fu, J., Liu, J., Tian, H., Li, Y., Bao, Y., Fang, Z., Lu, H.: Dual attention network for scene segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)

[15] Hu, P., Perazzi, F., Heilbron, F.C., Wang, O., Lin, Z., Saenko, K., Sclaroff, S.: Realtime semantic segmentation with fast attention. IEEE Robotics and Automation Letters 6(1), 263–270 (2020)

[16] Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Transformers in vision: A survey. arXiv preprint arXiv:2101.01169 (2021)

[17] Li, C., Tong, Q., Liao, X., Si, W., Sun, Y., Wang, Q., Heng, P.A.: Attention based hierarchical aggregation network for 3d left atrial segmentation. In: Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges. pp. 255–264 (2019)

[18] Milletari, F., Navab, N., Ahmadi, S.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV). pp. 565–571 (2016)

[19] Nie, D., Gao, Y., Wang, L., Shen, D.: Asdnet: Attention based semi-supervised deep networks for medical image segmentation. In: Frangi, A., Fichtinger, G., Schnabel, J., Alberola-L´opez, C., Davatzikos, C. (eds.) MICCAI 2018. pp. 370–378. Lecture Notes in Computer Science, Springer Verlag (2018)

[20] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241 (2015)

[21] Roy, A.G., Navab, N., Wachinger, C.: Concurrent spatial and channel squeeze excitation in fully convolutional networks. In: MICCAI. vol. abs/1803.02579 (2018)

[22] Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., Rueckert, D.: Attention gated networks: Learning to leverage salient regions in medical images. Medical Image Analysis 53 (02 2019). https://doi.org/10.1016/j.media.2019.01.012

[23] Sinha, A., Dolz, J.: Multi-scale self-guided attention for medical image segmentation. IEEE Journal of Biomedical and Health Informatics pp. 1–1 (2020)

[24] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NeurIPS. pp. 5998–6008 (2017)

[25] Wang, H., Zhu, Y., Green, B., Adam, H., Yuille, A., Chen, L.C.: Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In: European Conference on Computer Vision. pp. 108–126 (2020)

[26] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen, “Kvasir-seg: A segmented polyp dataset,” in MMM, 2020.