JADAVPUR UNIVERSITY

MASTER DEGREE PROJECT

UNET AND ATTENTION NETWORK-BASED DEEP LEARNING TECHNIQUE FOR POLYP SEGMENTATION

A project report submitted in partial fulfillment of the requirements for the degree of Master of Computer Application

In

Computer Science and Engineering

by

TITIRSHA GHOSH
Class Roll No: 002110503012
Examination Roll No.: MCA2340043
Registration No.: 160118

Under the Guidance of
Dr. Debotosh Bhattacharjee
Professor
Department of Computer Science and Engineering
Jadavpur University Kolkata-700032

Department of Computer Science and Engineering
Faculty of Engineering and Technology
Kolkata -700032
May 30, 2023

Declaration of Originality & Compliance of Academic Ethics:

I, Titirsha Ghosh, declare that this project report titled, “UNET AND ATTENTION NETWORK BASED DEEP LEARNING TECHNIQUE FOR POLYP SEGMENTATION” contains original research work by the undersigned candidate as a part of her Master of Computer Application studies.All information in this document has been obtained and presented in accordance with academic rules and ethical conduct.I also declared that ,as required by these rules and conduct,I have fully cited and referenced all materials and results that are not original to this work

Name : TITIRSHA GHOSH

Examination Roll No.: MCA2340043

Registration No.: 160118

Class Roll No : 002110503012

Project Title : U-Net and Attention Network Based Deep Learning Technique for Polyp Segmentation

---

Signature with Date

To whom it may concern

I hereby recommend that project report entitled “UNET AND ATTENTION Network-Based Deep Learning Technique for Polyp Segmentation” has been satisfactorily completed by Titirsha Ghosh, Roll No. 002110503012, Examination Roll No.: MCA2340043, Registration No. 160118 of 2021-2023 under my guidance and supervision may be accepted as partial fulfillment of the requirements for the degree of Master of Computer Application in Computer Science and Engineering from the Department of Computer Science and Engineering, Jadavpur University for the academic session 2021-2023.

Prof. Debotosh Bhattacharjee
(Project Supervisor)
Department of Computer Science & Engineering
Jadavpur University

Prof. Nandini Mukhopadhyay
Head of The Department
Department of Computer Science & Engineering
Jadavpur University

Prof. Ardhendu Ghoshal
Dean
Department of Computer Science & Engineering
Jadavpur University

Certificate of Approval

This is to certify that the project entitled “UNET AND ATTENTION Network-Based Deep Learning Technique for Polyp Segmentation” is a bonafide record of work carried out by Titirsha Ghosh in partial fulfillment of the requirements for the award of the degree of Master of Computer Application from the Department of Computer Science and Engineering, Jadavpur University for the academic session 2021-2023. It is understood that by this approval, the undersigned does not necessarily endorse or approve any statement made, opinion expressed, or conclusion drawn therein but approve the project only for the purpose for which it has been submitted.

Examiner:

Signature of Examiner 1

Date:

Signature of Examiner 2

Date:

Jadavpur University

Abstract

Faculty of Engineering and Technology, Jadavpur University

Computer Science and Engineering

Master of Computer Application

UNET AND Attention Network-Based Deep Learning Technique for Polyp Segmentation

By Titirsha Ghosh

Colon cancer is one of the leading causes of cancer-related deaths worldwide, and early detection of precancerous polyps is crucial for effective treatment. Polyp segmentation has accomplished massive triumph over the years in supervised learning. However, obtaining many labeled datasets is commonly challenging in the medical domain. To solve this problem, we developed a fully automated pixel-wise polyp segmentation model using the U-Net. The U-Net architecture is trained on Kvasir-SEG and CVC-ClinicDB, two open-access datasets of gastrointestinal polyp images and corresponding segmentation masks, manually annotated and verified by an experienced gastroenterologist. The network is designed to predict a pixel-wise segmentation mask of the polyp region in the input image. We demonstrate that the U-Net model achieves high accuracy in polyp segmentation. This paper also developed an extension of the U-Net architecture for polyp segmentation that incorporates an attention mechanism. During the segmentation process, the attention mechanism is used to selectively highlight relevant regions of the colonoscopy image, such as the polyp region. We demonstrate that the Attention U-Net model achieves higher accuracy in polyp segmentation than the original U-Net. The attention mechanism also provides insights into the importance of different regions of the colonoscopy image for accurate polyp segmentation. We believe this approach can improve colon cancer diagnosis and treatment accuracy and efficiency.

Acknowledgements

I would like to extend my gratitude to the people who helped to bring this project report work to completion. First and foremost, I would like to express my sincere gratitude and appreciation to my project supervisor **Prof. Debotosh Bhattacharjee** for providing me with his valuable advice and guidance throughout the project work.

I want to express my sincere, heartfelt gratitude to Prof. Nandini Mukhopadhyay, Head of the Department of Computer Science and Engineering, Jadavpur University, and Prof. Ardhendu Ghoshal, Dean, Faculty of Engineering and Technology, Jadavpur University, for providing me all the facilities and for their support to the activities of this research.

I thank my parents, who have always supported and inspired me.

Last, but not the least, I would like to thank all my classmates and all respected teachers for their valuable suggestions and helpful discussions.

Name : Titirsha Ghosh
Class Roll No.: 002110503012
Examination Roll No.: MCA2340043
Registration No.: 160118
Department of Computer Science and Engineering Jadavpur University

CONTENTS

PAGE NO

CHAPTER 1:

INTRODUCTION

1.1 Background

The human gastrointestinal (GI) tract comprises different sections, including the large bowel. Several anomalies and diseases, such as colorectal cancer, can affect the large bowel. Colorectal cancer is the second most common cancer type among women and the third most common among men. Polyps are precursors to colorectal cancer and are found in nearly half of the individuals at age 50 having a screening colonoscopy, and are increasing with age. Colonoscopy is the gold standard for detecting and assessing these polyps with subsequent biopsy and removal of the polyps. Early disease detection has a huge impact on survival from colorectal cancer, and polyp detection is therefore important. In addition, several studies have shown that polyps are often overlooked during colonoscopies, with polyp miss rates of 14%-30% depending on the type and size of the polyps. Increasing the detection of polyps has been shown to decrease the risk of colorectal cancer. Thus, the automatic detection of more polyps at an early stage can play a crucial role in improving both prevention and survival from colorectal cancer.

Polyp segmentation is a computer vision technique that automatically identifies and locates polyps in medical images, such as colonoscopies. The goal of polyp segmentation is to accurately detect the boundary of the polyp region in the image, which can help doctors to diagnose and treat polyps more effectively. Polyp segmentation is challenging due to the variability in polyp appearance, size, shape, location, and the presence of noise and artifacts in medical images.

Artificial intelligence (AI) maneuvers in colonoscopy have achieved encouraging and promising results in recent years. Deep learning (DL) is among the widely accepted tools in the AI field and is also a subfield of machine learning (ML). It is a method of extracting class-specific important features by stacking multiple nonlinear and linear blocks in deep layers, and the information is transferred between them. CNNs (Convolutional Neural Networks) are primarily used for image segmentation. It is inspired by how the human brain processes visual information, where the neurons in the visual cortex respond to specific patterns and features in the input image.

In recent years, deep-learning-based techniques have achieved significant success in the computer vision domain, and interest in applying deep learning to endoscopic image segmentation has grown. In particular, encoder-decoder-based methods such as U-Net, UNet++,

SegNet, and fully convolutional networks (FCNs) have been commonly used for semantic segmentation.

1.2 Features of U-Net

Here are some reasons why the U-Net is a popular choice for image segmentation:

1. Skip connections: The U-Net architecture incorporates skip connections that connect the corresponding layers in the encoder and decoder networks. These skip connections allow the U-Net to capture local and global information in the input image, which is important for accurate segmentation.
2. Fewer parameters: The U-Net has fewer parameters than other segmentation models, which makes it easier to train and less prone to overfitting.
3. Prevalence: The U-Net has been successfully applied to various image segmentation tasks, including biomedical image segmentation, cell segmentation, and road segmentation.
4. Versatility: The U-Net has been extended to other variants, such as the Attention U-Net and the V-Net, which incorporate attention mechanisms and 3D convolutions, respectively. This makes the U-Net a versatile architecture that can be adapted to different image segmentation tasks.

1.3 Features of Attention U-Net

The Attention U-Net is an extension of the U-Net architecture that incorporates attention mechanisms to improve the segmentation performance. Here are some reasons why the Attention U-Net is a popular choice for image segmentation:

1. Attention mechanisms: The Attention U-Net incorporates attention mechanisms that allow the model to focus on the most informative regions of the input image. This helps the model to segment the image more accurately and efficiently.
2. Better performance: Several studies have shown that the Attention U-Net outperforms the standard U-Net on various segmentation tasks, such as biomedical image segmentation and road segmentation.
3. Adaptive feature selection: The attention mechanisms in the Attention U-Net allow the model to adaptively select and weigh the most relevant features from the input image.

This makes the model more robust to variations in the input image and improves the segmentation accuracy.

4. Versatility: The Attention U-Net can be applied to various image segmentation tasks and easily integrated with other deep learning architectures.

1.4 Summary

In summary, this study makes the following contributions:

- We used a U-Net architecture for image segmentation tasks, trained the model using labeled data and Incorporated skip connections to connect corresponding layers in the encoder and decoder networks.
- We utilized an attention mechanism to selectively focus on the most informative features in the input image, improved accuracy in the segmentation results and made the model more robust to noise and artifacts in the input image.
- We evaluated our model on the Kvasir-SEG and CVC-ClinicDB datasets, and the experimental shows the result of intersection over union (IoU), Dice coefficient, recall, precision, and F1-score.

CHAPTER 2:

RELATED WORK

CNN-based polyp segmentation is a promising technique for automating polyp detection and segmentation in colonoscopy images. By leveraging the power of deep learning, CNN-based models can learn to identify and segment polyps in medical images with high accuracy and efficiency. These models can potentially improve the sensitivity and specificity of polyp detection, reduce inter-observer variability, and increase the overall quality of colon cancer screening and diagnosis. With further research and development, CNN-based polyp segmentation may become an important tool for improving the detection and treatment of colon cancer, ultimately leading to better patient outcomes and reduced healthcare costs.

There have been several studies on CNN-based polyp segmentation in recent years, with many demonstrating promising results. For example, a study published in the Journal of Biomedical Optics 2020 used a U-Net model to segment polyps in colonoscopy images. The authors reported an average Dice similarity coefficient of 0.836 and an average Intersection over Union (IoU) of 0.729, indicating high accuracy in the segmentation results.

Another study published in Computerized Medical Imaging and Graphics in 2021 used an Attention U-Net model to segment polyps in colonoscopy images. The authors reported an average IoU of 0.841 and an average precision of 0.829, indicating improved accuracy compared to previous methods.

Other studies have explored transfer learning, data augmentation, and other techniques to improve the performance of CNN-based polyp segmentation models. These studies demonstrate the potential of CNN-based polyp segmentation to improve the accuracy and efficiency of colon cancer screening and diagnosis and suggest that further research in this area is warranted.

CHAPTER 3:

STRUCTURE AND METHOD EXPLANATION

3.1 Overview

The network architecture is illustrated in Figure (a). It consists of a contracting path (left side) and an expansive path (right side). The contracting path follows the typical architecture of a convolutional network. It consists of the repeated application of two 3x3 convolutions (unpadded convolutions), each followed by a rectified linear unit (ReLU) and a 2x2 max pooling operation with stride 2 for downsampling. At each downsampling step, we double the number of feature channels. Every step in the expansive path consists of an upsampling of the feature map followed by a 2x2 convolution ("up-convolution") that halves the number of feature channels, a concatenation with the correspondingly cropped feature map from the contracting path, and two 3x3 convolutions, each followed by a ReLU. The cropping is necessary due to the loss of border pixels in every convolution. At the final layer, a 1x1 convolution maps each 16- component feature vector to the desired number of classes. In total, the network has 23 convolutional layers. It is important to select the input tile size so that all 2x2 max-pooling operations are applied to a layer with an even x- and y-size to allow a seamless tiling of the output segmentation map.

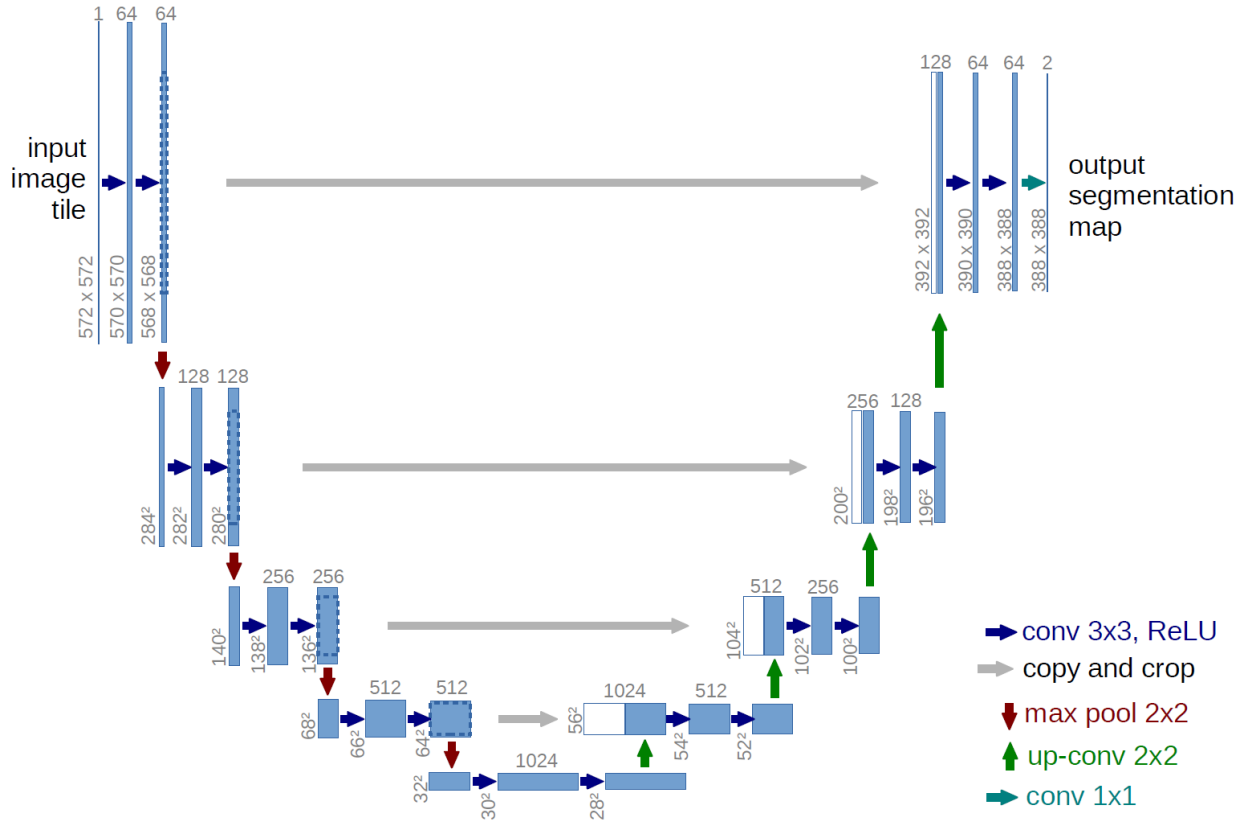
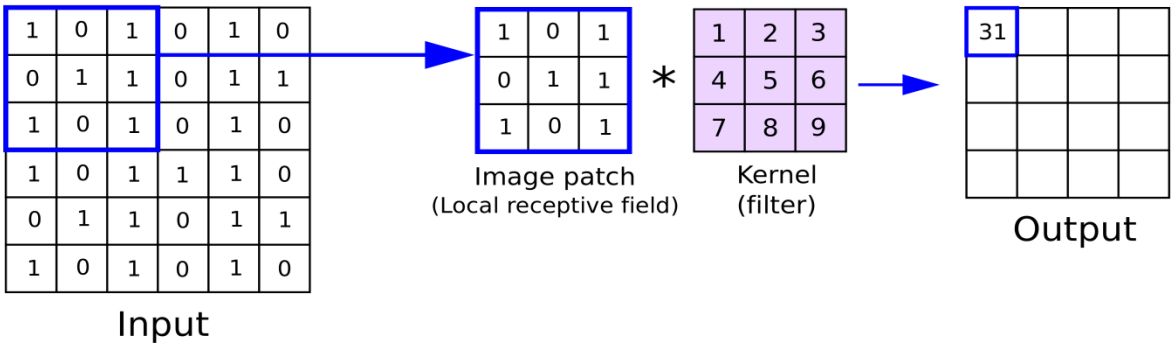


Fig (a)

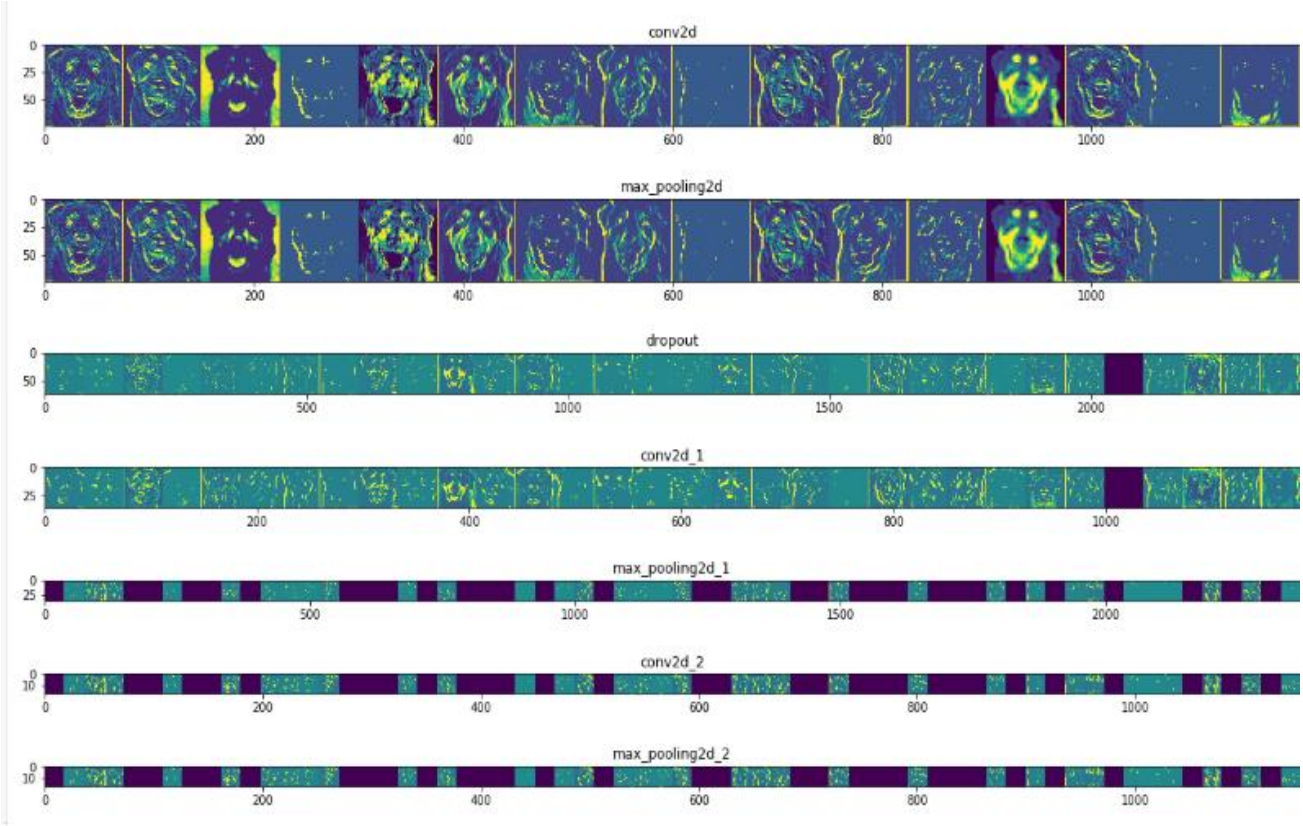
The mentioned model takes a training dataset X that consists of N sample images x: X = x1, x2, . . . , xN, with corresponding Y = y1, y2, . . . , yN. Then, each ground truth pixel i of any given sample y is y ∈ [0, 1]. For the Kvasir-SEG dataset, we feed our network with a 256 × 256 × 3 image and obtain a 256 × 256 × 1 output segmentation mask. For the CVC-ClinicDB dataset, also we feed our network with a 256 × 256 × 3 image and obtain a 256 × 256 × 1 output segmentation mask.

CNN:

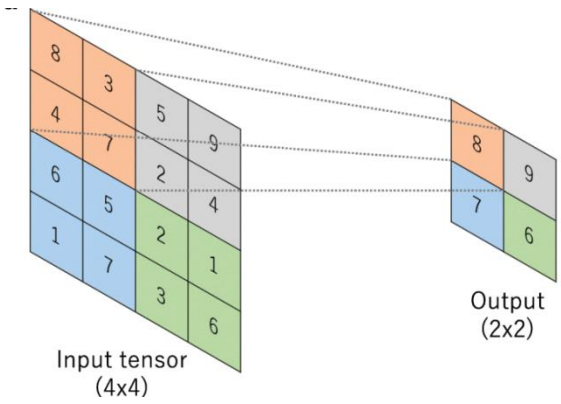
Convolution neural network is a technique that tries extracting features from images using filters and then mapping these features maps to a class or a label, Instead of naive DNN or deep neural network that maps the simple pixels with the class after a deep network of dense layers. And CNN technique shows great progress compared to naive DNN; here is how it works.



A filter of size (f,f) will go on its size of the image, then element-wise multiplication, then summation and put the sum as the first pixel of output and moving to the next image patch by moving step called stride (S), some times we use what is calling padding (p) to preserve the output shape as input shape. In this brilliant way, we can extract features like vertical or horizontal lines or even circles from an image like this.

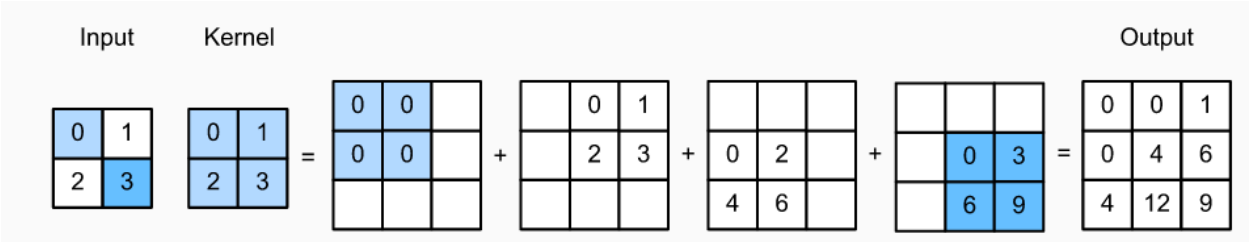


In the **encoding path**, an input image passes through convolutional layers with 3x3 filters and ReLU activation functions. These layers extract local features from the image and produce feature maps with the same spatial dimensions as the input. To avoid overfitting, dropout is applied after each convolutional layer. Dropout randomly sets a fraction of the input values to zero during training. After each pair of convolutional layers, a 2x2 max pooling layer is applied to the feature maps.



Max pooling reduces the spatial dimensions of the feature maps by a factor of two and retains the strongest activation within each pooling region. The same convolutional, dropout, and pooling layers sequence is repeated several times to extract increasingly complex features from the input image.

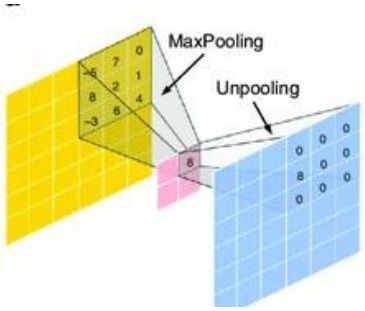
The **decoding path** of the U-Net model has the purpose of upsampling the low-resolution feature maps generated by the encoder to recover the original image resolution. It also combines the high-resolution feature maps obtained in the encoder with the low-resolution feature maps in the decoder to preserve spatial information.



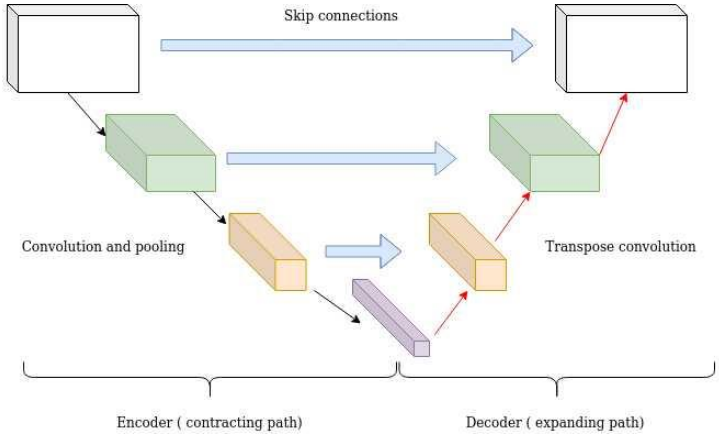
Here each input pixel is multiplied with the kernel and put the output(the same as the kernel size) in the final output feature map and moves again with a stride (but this time stride to the output). So, increasing the stride will increase the output size; conversely, increasing the padding will decrease the output size. Here is the output size equation

Output size = (input size-1)*stride -2*padding +(kernel size -1) +1

The convolutional transpose layer performs an inverse operation to the pooling operation in the encoding part, where the size of the feature maps is increased by a factor of two. The convolutional transpose operation also introduces learnable parameters allowing the network to upsample effectively. The feature maps obtained from the corresponding encoding part are concatenated with the upsampled feature maps. The purpose of concatenation is to combine the high-level information learned by the encoding part with the low-level information obtained in the decoding part. After concatenation, the combined feature maps are fed into several convolutional layers. These layers perform a series of convolutions with a kernel size 3x3, and apply the ReLU activation function. These layers help to refine the feature maps and recover the finer details of the image. Dropout layers are inserted between the convolutional layers to prevent overfitting and improve generalization. The final convolutional layer has one filter and a kernel size of 1x1. It outputs the probability map of the segmentation mask, which is fed into the loss function to optimize the parameters of the network.



Skip connections are used in the U-Net architecture to help address the problem of information loss during the downsampling (encoding) process. In traditional encoder-decoder networks, the input image is repeatedly downsampled, which can lead to the loss of important spatial information. The skip connections allow the model to retain this information by passing it from the encoder to the decoder, which can be used to reconstruct the original input.



After the u9 layer is created and two more convolutional layers with 16 filters of size 3x3 and a ReLU activation function are applied to the concatenated feature maps, we use 1 × 1 convolution and sigmoid activation, as shown in Equation (1), to obtain the final segmentation map.

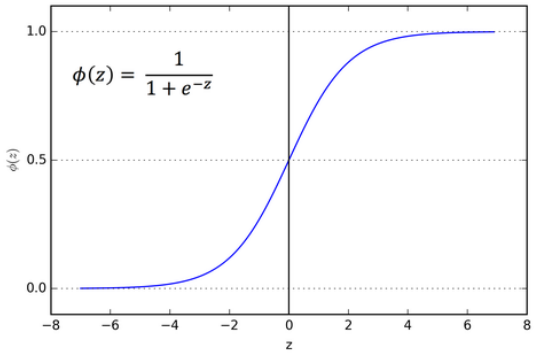
A loss function is a function that compares the target and predicted output values; measures how well the neural network models the training data. We aim to minimize this loss between the predicted and target outputs when training.

The hyperparameters are adjusted to minimize the average loss — we find the weights, *wT*, and biases, *b*, that minimize the value of *J* (average loss).

$$J(w^T, b) = \frac{1}{m}\sum_{i=1}^{m} L(\hat{y}^{(i)}, y^{(i)})$$

Here we trained our network with the **Dice Coefficient Loss function** based on the ground truth for the training images. Equation (2) shows the formula of the loss function, where L is the loss for a prediction yi consisting of N pixels at a specific output of the network:

Sigmoid function, $y=1 / (1+e^{-x})$ (1)



Dice Loss (L) = 1- Dice similarity Score

= 1-(2 * Area of Overlap+1)/(total pixels combined+1) (2)

For the case of evaluating a Dice coefficient on predicted segmentation masks, we can approximate the Area of Overlap as the element-wise multiplication between the prediction and target mask and then sum the resulting matrix.

$$|A \cap B| = \begin{bmatrix} 0.01 & 0.03 & 0.02 & 0.02 \\ 0.05 & 0.12 & 0.09 & 0.07 \\ 0.89 & 0.85 & 0.88 & 0.91 \\ 0.99 & 0.97 & 0.95 & 0.97 \end{bmatrix} * \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \xrightarrow{\text{element-wise multiply}} \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0.89 & 0.85 & 0.88 & 0.91 \\ 0.99 & 0.97 & 0.95 & 0.97 \end{bmatrix} \xrightarrow{\text{sum}} 7.41$$

prediction target

3.2 Details of each layer in U-Net

Table 1:

	Feature Size	
input	256×256×1	
convolution 1	256×256×16	[3 × 3,16, stride 1] × 2
Pooling 1	128×128×16	2 × 2 max pool, stride 2
convolution 2	128×128×32	[3 × 3,32, stride 1] × 2
Pooling 2	64×64××32	2 × 2 max pool, stride 2
convolution 3	64×64××64	[3 × 3,64, stride 1] × 2
Pooling 3	32×32×64	2 × 2 max pool, stride 2
convolution 4	32×32×128	[3 × 3,128, stride 1] × 2
Pooling 4	16×16×128	2 × 2 max pool, stride 2
convolution 5	16×16×256	[3 × 3,256, stride 1] × 2
Upsample 6	32×32×256	2 × 2 upsample, stride 2
convolution 6	32×32×128	[3 × 3,128, stride 1] × 2
Upsample 7	64×64×128	2 × 2 upsample, stride 2
convolution 7	64×64×64	[3 × 3,64, stride 1] × 2
Upsample 8	128×128×64	2 × 2 upsample, stride 2
convolution 8	128×128×32	[3 × 3,32, stride 1] × 2
Upsample 9	256×256×32	2 × 2 upsample, stride 2
convolution 9	256×256×16	[3 × 3,16, stride 1] × 2
Final Output	256×256×1	[1 × 1,16, stride 1]

Note that [3 × 3,16, stride 1] corresponds to 3 × 3 kernel size convolution with 16 features with stride 1. "[] × n" indicates n iterations of the convolution block.

3.3 Attention U-Net

Over the last few years, the attention mechanism has become very popular in various deep learning research areas, starting with natural language processing (NLP). Recently, it has been applied to computer vision tasks. The attention model has been utilized as a pixel-wise prediction

model in the semantic segmentation domain. It identifies the sections of the network that need more attention. Continuous use of an attention mechanism at each level allows long-range spatial dependency of feature maps. The attention block also decreases each image's computing cost to a fixed dimensional vector. Therefore, the fundamental value of an attention unit is that it is straightforward and can be applied to every input scale to strengthen the consistency of the features that emphasize the result.

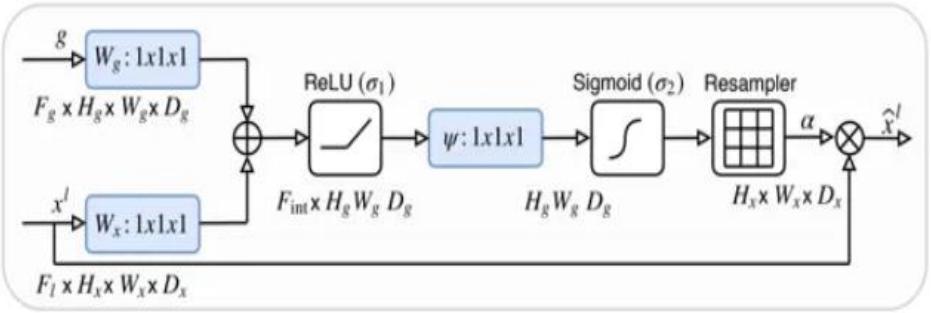
ATTENTION: In the context of image segmentation, attention is a way to highlight only the relevant activations during training. This reduces the computational resources wasted on irrelevant activations, giving the network better generalization power. The network can pay "attention" to certain parts of the image. Attention comes in two forms, hard and soft.

- Hard Attention:

Hard attention works based on highlighting relevant regions by cropping the image or iterative region proposal. Since hard attention can only choose one region of an image at a time, it has two implications, it is non-differentiable and requires reinforcement learning to train. Since it is non-differentiable, it means that for a given region in an image, the network can either pay "attention" or not, with no in-between.

- Soft Attention:

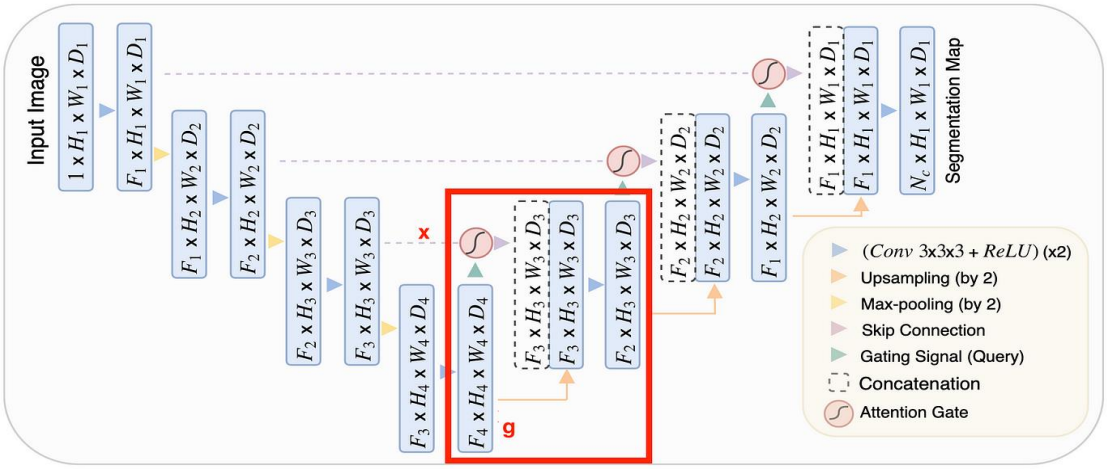
Soft attention works by weighing different parts of the image. Areas of high relevance are multiplied with a larger weight, and areas of low relevance are tagged with smaller weights. As the model is trained, more focus is given to the regions with higher weights. Unlike hard attention, these weights can be applied to many patches in the image.
The attention gates introduced by Oktay et al. use additive soft attention.



pic: Attention-Block

The attention gate takes in two inputs, vectors ***x*** and ***g***.

- The vector, ***g***, is taken from the next lowest layer of the network. The vector has smaller dimensions and better feature representation because it comes deeper into the network.
- Vector ***x*** would have dimensions of 128x128x128(filters x height x width), and vector ***g*** would be 64x64x64.
- Vector ***x*** goes through a stridden convolution such that its dimensions become 64x64x128, and vector ***g*** goes through a 1x1 convolution such that its dimensions become 64x64x128.
- The two vectors are summed element-wise. This process results in aligned weights becoming larger while unaligned weights become relatively smaller.
- The resultant vector goes through a ReLU activation layer and a 1x1 convolution that collapses the dimensions to 64x64x1.
- This vector goes through a sigmoid layer which scales the vector between the range [0,1], producing the attention coefficients (weights), where coefficients closer to 1 indicate more relevant features.
- The attention coefficients are upsampled to the original dimensions (128x128) of the ***x*** vector using trilinear interpolation. The attention coefficients are multiplied element-wise to the original ***x*** vector, scaling the vector according to relevance. This is then passed along in the skip connection as normal.



Pic: Attention-Unet

CHAPTER 4:

Experiments

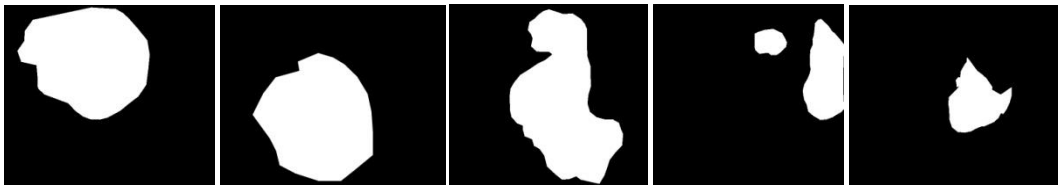
We evaluated our discussed model, i.e., U-Net & Attention U-Net architecture on public segmentation datasets, Kvasir-SEG & CVC-ClinicDB.

4.1 Datasets

Kvasir-SEG Dataset:

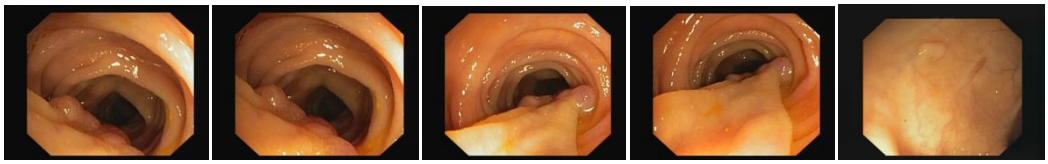
The Kvasir-SEG dataset (size 46.2 MB) contains 1000 polyp images and their corresponding ground truth from the Kvasir Dataset v2. The images' resolution in Kvasir-SEG varies from 332x487 to 1920x1072 pixels. The images and their corresponding masks are stored in two separate folders with the same filename. The image files are encoded using JPG compression, facilitating online browsing. Training and testing were performed with an image resolution of 256×256 pixels. The images were randomly split into 80% for training and 20% for testing and validation.

Polyp Image 

Ground Truth 

CVC-ClinicDB Dataset:

CVC-ClinicDB is an open-access dataset of 612 images with a resolution of 384×288 from 31 colonoscopy sequences. It is used for medical image segmentation, particularly polyp detection in colonoscopy videos. CVC-ClinicDB is a database of frames extracted from colonoscopy videos. The image files are encoded using PNG compression, facilitating online browsing. Training and testing were performed with an image resolution of 256 × 256 pixels. The images were randomly split into 80% for training and 20% for testing and validation.

Polyp Image 

Ground Truth 

4.2 Evaluation Metrics

Different metrics for evaluating and comparing the performance of the architectures exist. The most commonly used metrics for medical image segmentation tasks are the Dice coefficient and IoU. These are used in particular for several medical-related Kaggle competitions. In this medical image segmentation approach, each image pixel belongs to a polyp or non-polyp region. We calculate the Dice coefficient and mean IoU based on this principle.

Dice coefficient: The dice coefficient is a standard metric for comparing the pixel-wise results between predicted segmentation and ground truth. It is defined as:

$$\text{Dice coefficient } (A, B) = (2 \times |A \cap B|) /(|A| + |B)= 2 \times T P /(2 \times T P + F P + F N) \quad (1)$$

Where A signifies the predicted set of pixels and B is the ground truth of the object to be found in the image. Here, TP represents the true positive, FP represents the false positive, and FN represents the false negative.

Intersection over Union: The Intersection over Union (IoU) is another standard metric to evaluate a segmentation method. The IoU calculates the similarity between predicted (A) and its corresponding ground truth (B) as shown in the equation below:

$$IoU(A, B) = A \cap B /A \cup B = T P(t) /[T P(t) + F P(t) + F N(t)] \quad (2)$$

In equation 2, t is the threshold. At each threshold value t, a precision value is calculated based on the above equation and parameters, which is done by calculating the predicted object to all the ground truth objects.
To evaluate the polyp segmentation, we also used the following well-known segmentation evaluation metrics: recall, precision, and f1-score.

$$\textbf{Recall} = T P /(T P + F N)$$
$$\textbf{Precision} = T P /(T P + F P)$$
$$\textbf{F1-score}=2* \text{ Precision}* \text{ Recall}/(\text{Recall}+ \text{Precision})$$

4.3 Implementation Details

For **U-Net**, we trained all the methods in the Keras framework with TensorFlow as a backend.

We trained our model with 256×256-pixel images for the KVASIR-SEG dataset. We set the batch size to 16 and trained the model for 80 epochs. We used the Adam optimizer. We chose ReLU as the non-linear activation and Dice-coefficient loss as the loss function. To convert the predicted pixels to the background or foreground, we used a threshold value of 0.5.

We trained our model with 256 × 256-pixel images for the CVC-ClinicDB dataset. We set the batch size to 16 and trained the model for 100 epochs. We used the Adam optimizer. We chose ReLU as the non-linear activation and Dice-coefficient loss as the loss function. To convert the predicted pixels to the background or foreground, we used a threshold value of 0.32.

For **Attention-UNet**, we also trained all the Keras framework methods with TensorFlow as a backend.

We trained our model with 256×256-pixel images for the KVASIR-SEG dataset. We set the batch size to 16 and trained the model for 130 epochs. We used the Adam optimizer with a learning rate of 1e-3. We chose ReLU as the non-linear activation and Dice-coefficient loss as the loss function. To convert the predicted pixels to the background or foreground, we used a threshold value of 0.5.

We trained our model with 256 × 256-pixel images for the CVC-ClinicDB dataset. We set the batch size to 16 and trained the model for 130 epochs. We used the Adam optimizer with a learning rate of 1e-3. We chose ReLU as the non-linear activation and Dice-coefficient loss as the loss function. To convert the predicted pixels to the background or foreground, we used a threshold value of 0.4.

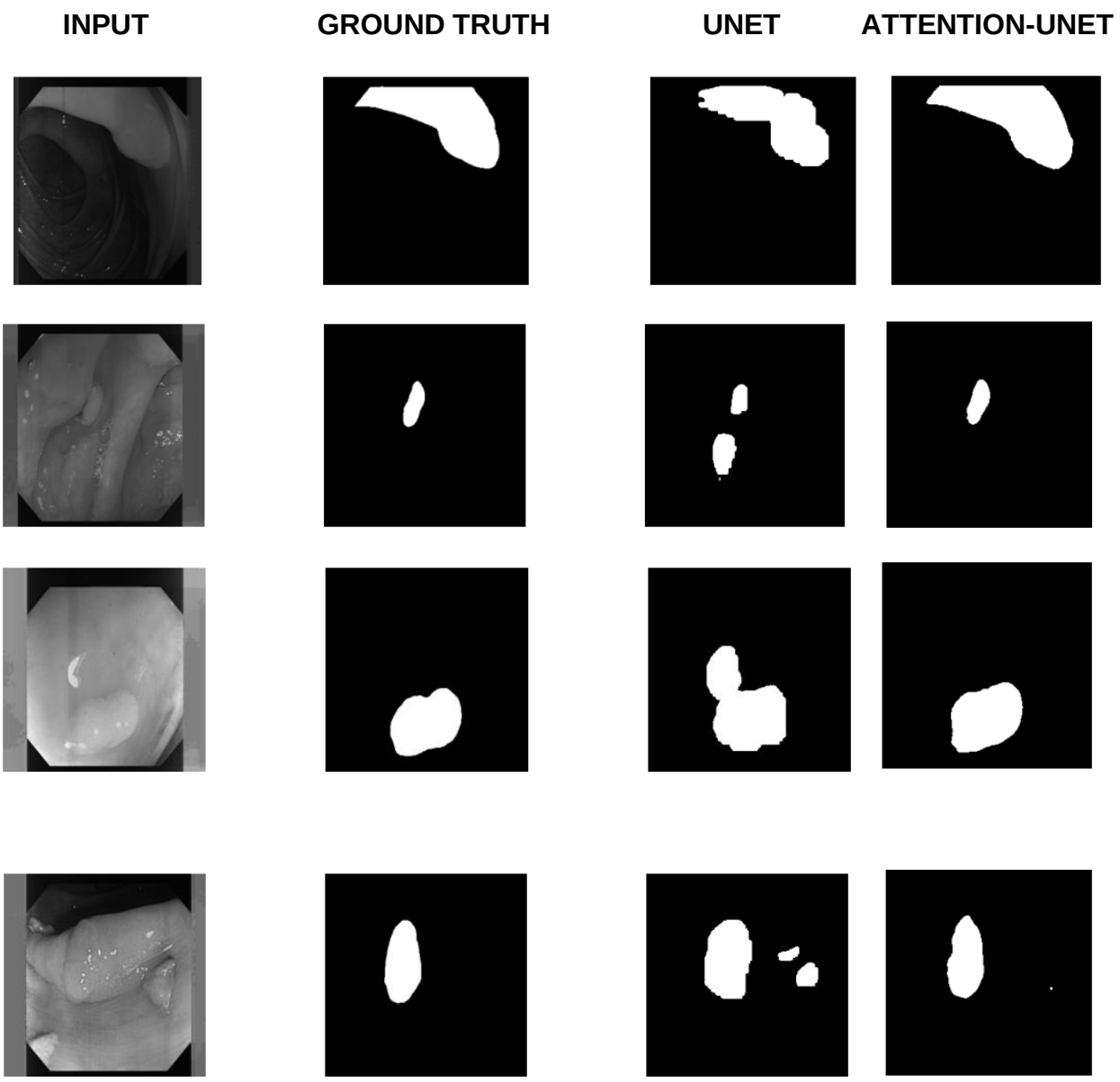
CHAPTER 5:

Results

5.1 Output Images

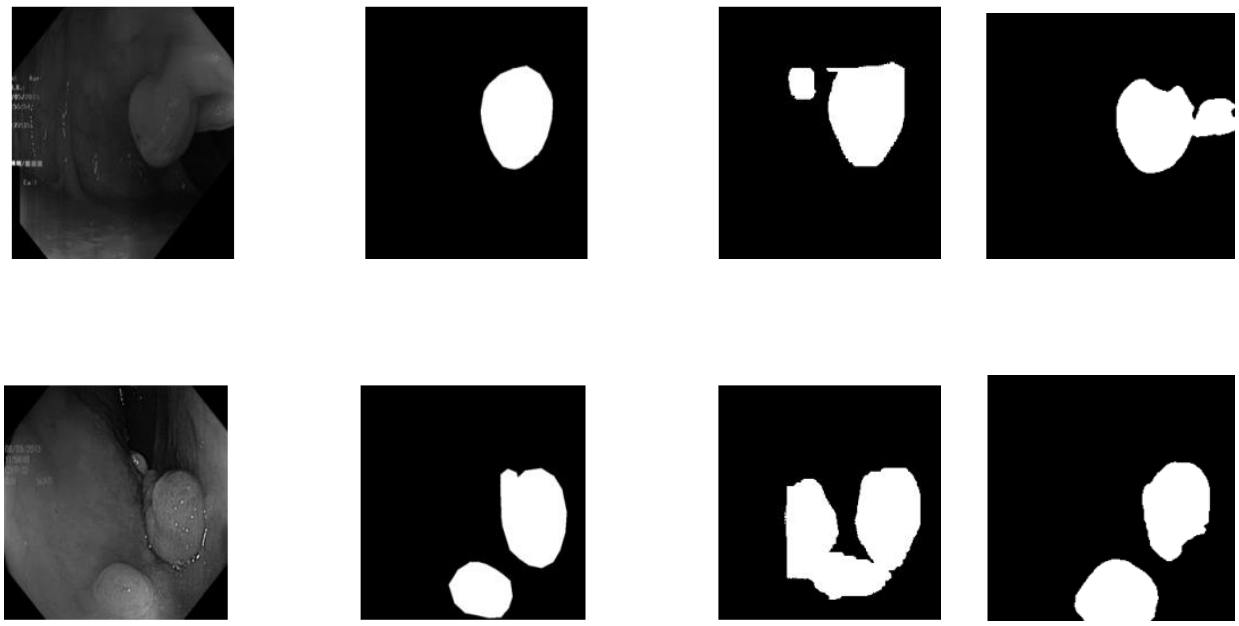
Here is the comparison of those evaluation metrics for UNet & Attention-UNet architecture:

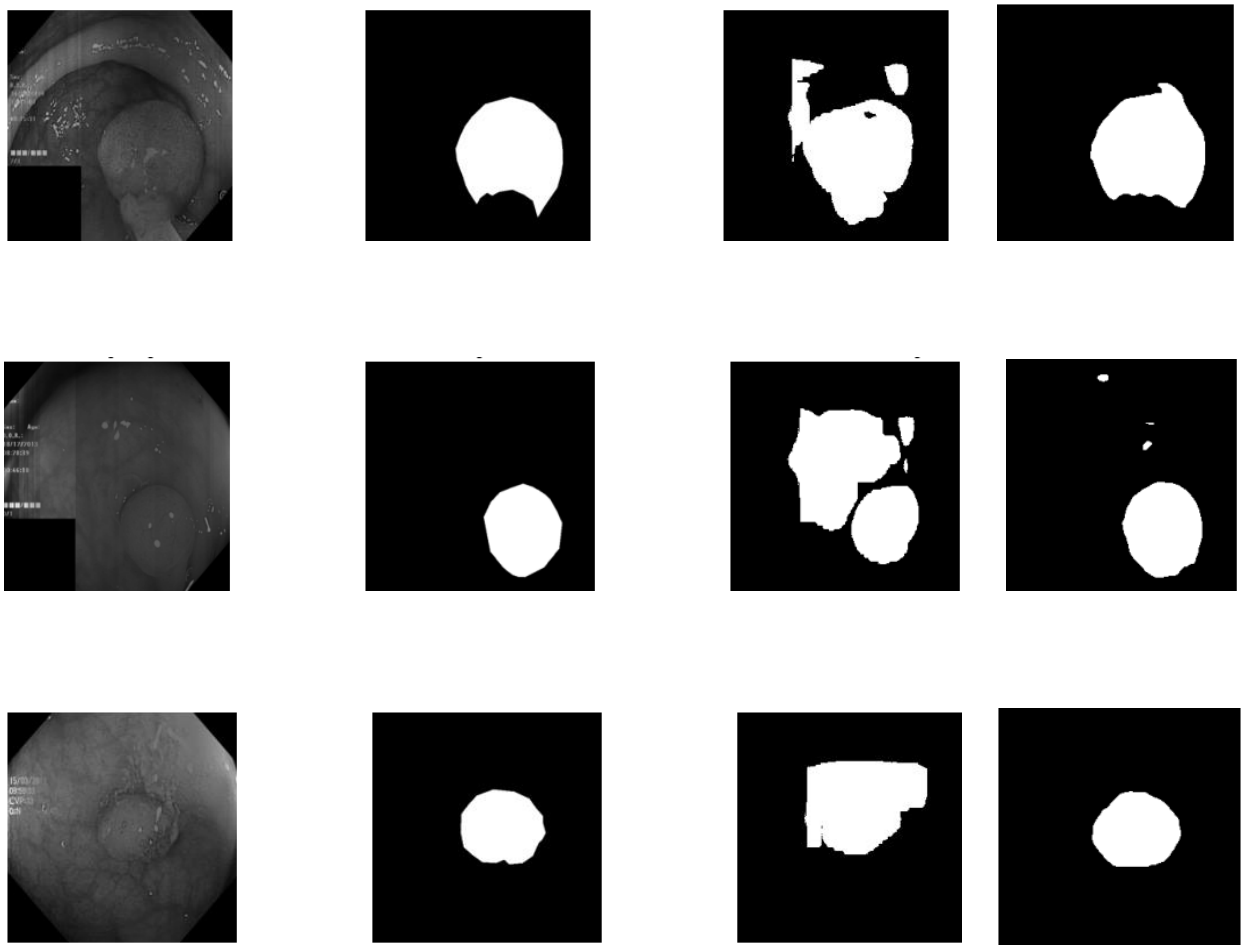
FOR CVC-ClinicDB:



FOR Kvasir-SEG:

INPUT	GROUND TRUTH	UNET	ATTENTION-UNET

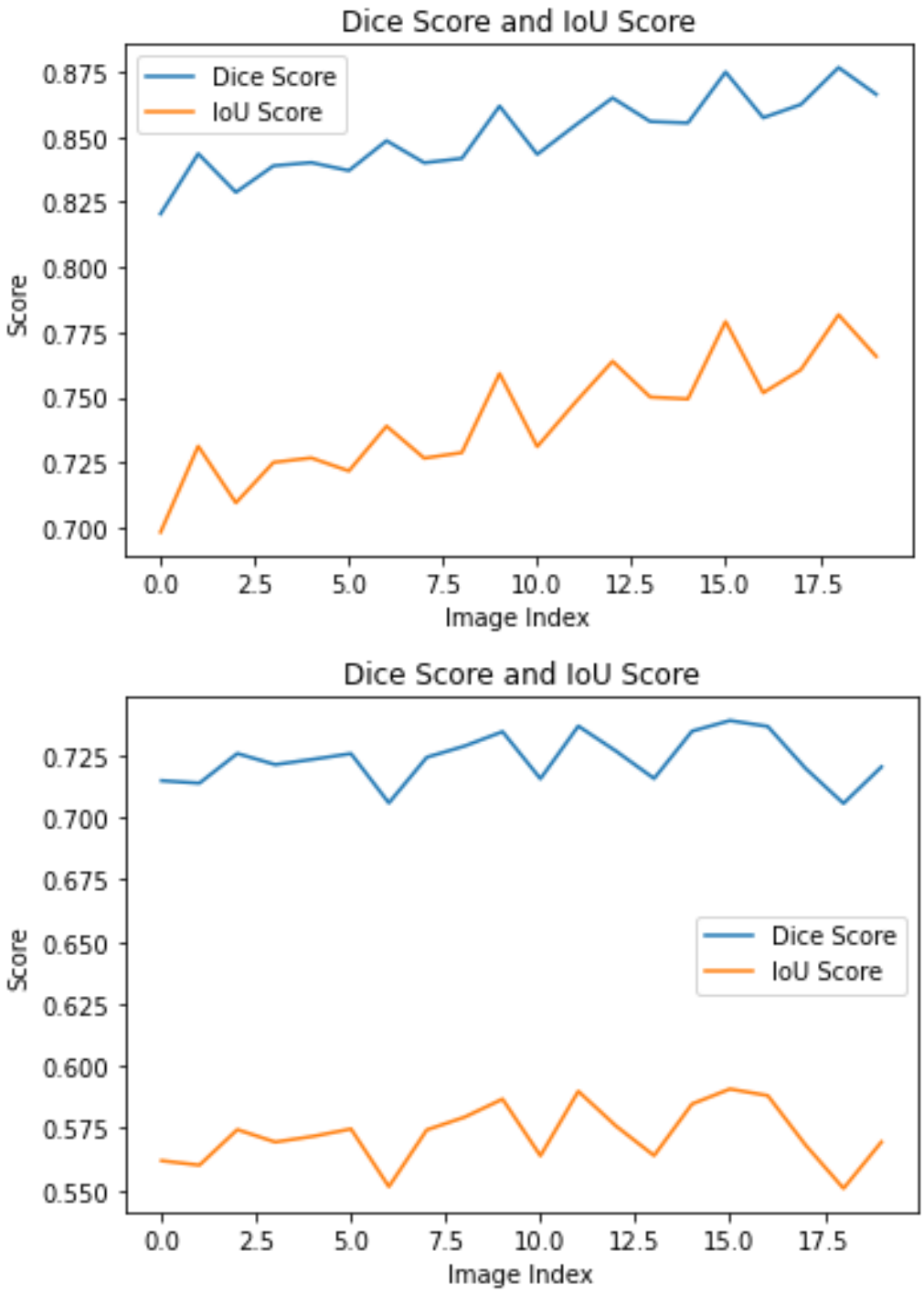




5.2 Comparison of metrics for U-NET and Attention U-NET Model

METHOD	DATASET	DICE-SCORE	IoU-SCORE	RECALL	PRECISION	F1-SCORE
UNET	CVC-ClinicDB	0.79	0.65	0.73	0.86	0.79
	Kvasir-SEG	0.5	0.32	0.61	0.43	0.5
ATTENTION-UNET	CVC-ClinicDB	0.86	0.78	0.95	0.98	0.96
	Kvasir-SEG	0.92	0.85	0.94	0.99	0.97

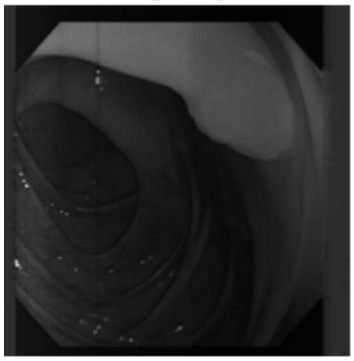
5.3 Comparison between train and test set w.r.t Dice score and IoU score:



This shows that training images have better scores than test images.

5.4 ANALYSIS OF OUTPUT WITH RESPECT TO THE NUMBER OF EPOCH

INPUT



GROUND TRUTH

AFTER 30 EPOCH

AFTER 60 EPOCH

AFTER 100 EPOCH

AFTER 130 EPOCH

CHAPTER 6:

Discussion

The UNet and Attention UNet models achieved results on Kvasir-SEG and CVC-ClinicDB datasets. From the qualitative results, it is obvious that the Attention UNet model's segmentation mask performed better than another method to capture the shape of information on the Kvasir-SEG dataset & CVC-ClinicDB dataset.

During the training process, we used various loss functions to improve our results, such as Jaccard loss, Dice loss, mean square loss, and binary cross-entropy loss. According to our experiments, the UNet & the Attention UNet method achieved a better result with a Dice-coefficient loss as loss function. We chose the loss function based on our analytical assessment. In addition, we found that the number of kernels, batch size, optimizer, loss function, and depth of the model may affect the result.

CHAPTER 7:

Conclusions

This paper presents biomedical image segmentation architecture, Unet, and later Attention UNet to achieve more accurate segmentation results. One of the limitations of the UNET architecture is that it may overlook fine-grained details in the input image due to the downsampling process. To address this issue, researchers have proposed different extensions of the UNET architecture, such as Attention UNET, which incorporates attention mechanisms to capture important regions of the input image better.

The Attention UNET architecture was proposed in 2018 by Oktay et al. It includes an attention mechanism that uses a gating network to learn how much attention to give to each feature map. The gating network takes as input the feature maps and a query vector, which is learned from the feature maps using a convolutional layer. The output of the gating network is a set of attention maps, which are multiplied element-wise with the feature maps to highlight important regions.

The Attention UNET architecture has achieved state-of-the-art results on several medical image segmentation tasks, such as brain tumor and liver segmentation. It is particularly useful in cases where the input images have complex and diverse textures, as it can adaptively learn to focus on the most informative regions of the input image.

To evaluate our approach's effectiveness, we conduct experiments on the Kvasir-SEG, CVC-ClinicDB dataset. The qualitative results show that the Attention UNet model’s segmentation mask performed better than other methods.

REFERENCES

1. Silva, J.; Histace, A.; Romain, O.; Dray, X.; Granado, B. Toward embedded detection of polyps in wce images for early diagnostics of colorectal cancer. Int. J. Comput. Assist. Radiol. Surg. 2014, 9, 283–293.

2. Arnold, M.; Sierra, M.S.; Laversanne, M.; Soerjomataram, I.; Jamel, A.; Bray, F. Global patterns and trends in colorectal cancer incidence and morality. Gut 2016, 66, 683–691.

3. Leufkens, A.M.; van Oijen, M.G.H.; Vleggaar, F.P.; Siersema, P.D. Factors influencing the miss rate of polyps in a back-to-back colonoscopy study. Endoscopy 2012, 44, 470–475.

4. Rabeneck, L.; El-Serag, H.B.; Davila, J.A.; Sandler, R.S. Outcomes of colorectal cancer in the United States: No change in survival (1986–1997). Am. J. Gastroenterol. 2003, 98, 471–477.

5. Mori, Y.; Kudo, S.-E. Detecting colorectal polyps via machine learning. Nat. Biomed. Eng. 2018, 2, 713.

6. Jha, D.; Smedsrud, P.H.; Riegler, M.A.; Johansen, D.; de Lange, T.; Halvorsen, P.; Johansen, H.D. ResUNet++: An advanced architecture for medical image segmentation. In Proceeding of IEEE International Symposium on Multimedia (ISM), San Diego, CA, USA, 9–11 December 2019; pp. 225–2255.

7. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. 2015, 115, 211–252.

8. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv 2014, arXiv:1409.1556.

9. Yun, K.; Park, J.; Cho, J. Robust human pose estimation for rotation via self-supervised learning. IEEE Access 2020, 8, 32502–32517.

10. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image sementation. In Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.

11. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. Unet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation. IEEE Trans. Med. Imaging. 2020, 39, 1856–1867.

12. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 2017, 39, 2481–2495.

13. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.

14. Casella, A.; Moccia, S.; Frontoni, E.; Paladini, D.; Momi, E.D.; Mattos, L.S. Inter-foetus membrane segmentation for TTTS using adversarial networks. Ann. Biomed. Eng. 2020, 48, 848–859. [CrossRef]

15. Li, W.; Liu, K.; Zhang, L.; Cheng, F. Object detection based on an adaptive attention mechanism. Sci. Rep. 2020, 10, 1–13.

16. Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; Tang, X. Residual attention network for image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3156–3164.

17. Li, K.; Wu, Z.; Peng, K.-C.; Ernst, J.; Fu, Y. Tell me where to look: Guided attention inference network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 9215–9223.

18. Chen, L.-C.; Yang, Y.; Wang, J.; Xu, W.; Yuille, A.L. Attention to scale: Scale-aware semantic image segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3640–3649.

19. Zhao, H.; Zhang, Y.; Liu, S.; Shi, J.; Loy, C.C.; Lin, D.; Jia, J. Psanet: Point-wise spatial attention network for scene parsing. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 267–283.

20. 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 Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154.

21. Attanasio, A.; Alberti, C.; Scaglioni, B.; Marahrens, N.; Frangi, A.F.; Leonetti, M.; Biyani, C.S.; Momi, E.D.; Valdastri, P. A Comparative Study of Spatio-Temporal U-Nets for Tissue Segmentation in Surgical Robotics. IEEE Trans. Med Robot. Bionics 2021.

22. Karkanis, S.; Iakovidis, D.K.; Maroulis, D.E.; Karras, D.; Tzivras, M. Computer-aided tumor detection in endoscopic video using color wavelet features. Inf. Technol. Biomed. IEEE Trans. 2003, 7, 141–152.