Image Segmentation on Medical Imaging Data

KUSHILUV JANGU, ABHINN YADAV
2020076, 2020013

on May 09, 2023

BTP Track: Research

BTP Advisor
Dr. Anubha Gupta

Indraprastha Institute of Information Technology
New Delhi

Student's Declaration

I hereby declare that the work presented in the report entitled **"Image Segmentation on Medical Imaging Data"** submitted by us for the partial fulfillment of the requirements for the degree of *Bachelor of Technology* in *Computer Science* at Indraprastha Institute of Information Technology, Delhi, is an authentic record of our work carried out under guidance of **Dr. Anubha Gupta**. Due acknowledgements have been given in the report to all material used. This work has not been submitted anywhere else for the reward of any other degree.

Kushiluv Jangu, Abhinn Yadav
Place & Date: New Delhi, May 09, 2023

Certificate

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Dr. Anubha Gupta **Place & Date: New Delhi, May 09, 2023**

Abstract

In this study, we compared the performance of various semantic and instance segmentation models on two different datasets: SegPC-2021 and DataScienceBowl 2018. The models tested were UNet, UNet++, Segformer for semantic segmentation and Mask RCNN for instance segmentation. Our results indicate that Mask RCNN outperforms other models in terms of IoU scores.

Keywords: Image Segmentation, Deep Learning, UNet,

Acknowledgments

We wish to express our deepest gratitude and sincerest thanks to our project advisor, Dr. Anubha Gupta, for her continuous encouragement, guidance and motivation. Her insightful feedback has been instrumental in shaping the direction of this work. We also extend our appreciation to our parents who motivated us to the successful competition of this project.

Work Distribution

....

Contents

This page intentionally left blank.

Chapter 1

Introduction

Medical image segmentation is a critical task in the field of medical imaging and has been gaining significant attention in recent years due to its potential for improving patient care and diagnosis accuracy. Image segmentation is the process of dividing an image into multiple segments or regions based on certain characteristics or features. In medical imaging, segmentation is often used to extract and identify specific regions of interest, such as tumors, organs, or cells.

In this report, we focus on the application of image segmentation on medical imaging data using four different models: U-Net, U-Net++, Segformer for semantic segmentation, and Mask R-CNN for instance segmentation. Semantic segmentation refers to the process of assigning a label to each pixel in the image, while instance segmentation involves identifying and distinguishing individual objects within an image. We implemented these models on two datasets, the Data Science Bowl 2018 and SEGPC-2021 Segmentation of Multiple Myeloma Plasma Cells in Microscopic Images.

The results of our experiments indicate that these models can achieve high segmentation accuracy on both datasets, demonstrating their potential for improving medical imaging analysis and diagnosis. Overall, this report provides insights into the application of image segmentation in the field of medical imaging and highlights the potential of various state-of-the-art models for this task.

Chapter 2

Background

2.0.1 Definitions

- **Image Segmentation**: The process of dividing an image into multiple segments or regions based on certain characteristics or features, with the goal of extracting and identifying specific regions of interest within the image.
- **Semantic Segmentation**: A type of image segmentation that assigns a label to each pixel in the image based on the content of the image. This technique is used to identify and segment regions of interest such as organs or tumors in medical images.
- **Instance Segmentation**: A type of image segmentation that involves identifying and distinguishing individual objects within an image. This technique is used to detect and segment multiple instances of the same object, such as cells in medical images.
- **U-Net**: A popular convolutional neural network architecture for semantic segmentation. It has an encoder-decoder structure and skip connections between the encoder and decoder, allowing it to effectively capture both low-level and high-level features in the image.
- **U-Net++**: An extension of the U-Net architecture that includes multiple nested paths and skip connections, further improving the accuracy of semantic segmentation.
- **Segformer**: A transformer-based architecture for semantic segmentation that uses

self-attention mechanisms to capture long-range dependencies and improve segmentation accuracy.

- **Mask R-CNN**: A popular instance segmentation model that extends the Faster R-CNN object detection model by adding a segmentation mask head, which predicts a binary mask for each detected object instance.

- **Medical Imaging**: The use of various imaging technologies such as X-ray, CT, MRI, and ultrasound to visualize the internal structures and functions of the human body for diagnosis and treatment of medical conditions.

2.0.2 Motivation

Medical image segmentation is a critical task in the field of medical imaging, with the potential to improve diagnosis accuracy and patient care. Accurate segmentation can help identify specific regions of interest, such as tumors or organs, and aid in treatment planning and disease monitoring. However, manual segmentation by medical professionals is time-consuming, subjective, and prone to human error.

With the advent of deep learning techniques and the availability of large medical imaging datasets, automated image segmentation has become a popular area of research. Various deep learning models have been proposed for image segmentation tasks, including semantic and instance segmentation. These models have shown promising results in accurately segmenting medical images and have the potential to improve diagnosis accuracy and patient care.

The motivation behind this study is to explore the potential of state-of-the-art deep learning models for medical image segmentation tasks and evaluate their effectiveness on different datasets. Our goal is to contribute to the development of automated medical image analysis and improve patient care and diagnosis accuracy.

Chapter 3

Datasets

In this section, we discuss the two datasets used in this report: Data Science Bowl 2018 and SEGPC-2021 Segmentation of Multiple Myeloma Plasma Cells in Microscopic Images.

SEGPC-2021 Segmentation of Multiple Myeloma Plasma Cells in Microscopic Images

SEGPC-2021 is a competition that aims to automate the segmentation of plasma cells in bone marrow aspirate slides of patients diagnosed with Multiple Myeloma (MM), a type of white blood cancer. The dataset's main author is the professor we are working under , Dr. Anubha gupta. It consists of microscopic images of bone marrow aspirate slides stained using Jenner-Giemsa stain. The images were captured in raw BMP format using two cameras attached to a microscope. The dataset contains a training set, a validation set, and a final test set. The training set contains 298 images, and the validation set contains 200 images. The final test set is used for evaluation, and its ground truth is not provided. The images were provided in two different resolutions: 2040x1536 pixels and 1920x2560 pixels.

Data Science Bowl 2018

The dataset was obtained from Kaggle and consists of microscopic images of cells. The dataset consists of microscopic images of cells, and the task is to identify the cells' nuclei, which is the starting point for most analyses. Identifying nuclei helps researchers understand the underlying biological processes at work by measuring how cells react to various treatments. The dataset contains two sets of images: a training set and a test set. The training

set contains 670 images, and the test set contains 65 images. The images were provided in various formats, including TIFF, JPEG, and PNG.

In both datasets, manual segmentation is a time-consuming and error-prone process that requires expert knowledge. Therefore, automated segmentation methods are essential to improve efficiency and accuracy. In the next section, we discuss the methods used in this report for automated image segmentation on these datasets.

Chapter 4

Current implemented work

4.1 UNET

4.1.1 Model Details

The U-Net is a convolutional neural network (CNN) architecture that was originally developed for biomedical image segmentation. It is named after its U-shaped architecture, which has a contracting path on the left-hand side and an expansive path on the right-hand side. The contracting path is made up of convolutional and pooling layers, which capture the high-level features of the input image. The expansive path consists of convolutional and upsampling layers, which enable the network to reconstruct the image from the high-level features.

4.1.2 Model Architecture

The U-Net architecture consists of a contracting path and an expansive path, which are connected by a bottleneck layer. The contracting path is composed of convolutional and max pooling layers, while the expansive path consists of convolutional and upsampling layers. The skip connections between the contracting and expansive paths allow the network to preserve fine-grained details.

Let X be the input image, and Y be the segmentation mask. The contracting path takes X as input and outputs a set of feature maps F_i, where i represents the depth of the network. The ith layer of the contracting path can be defined as:

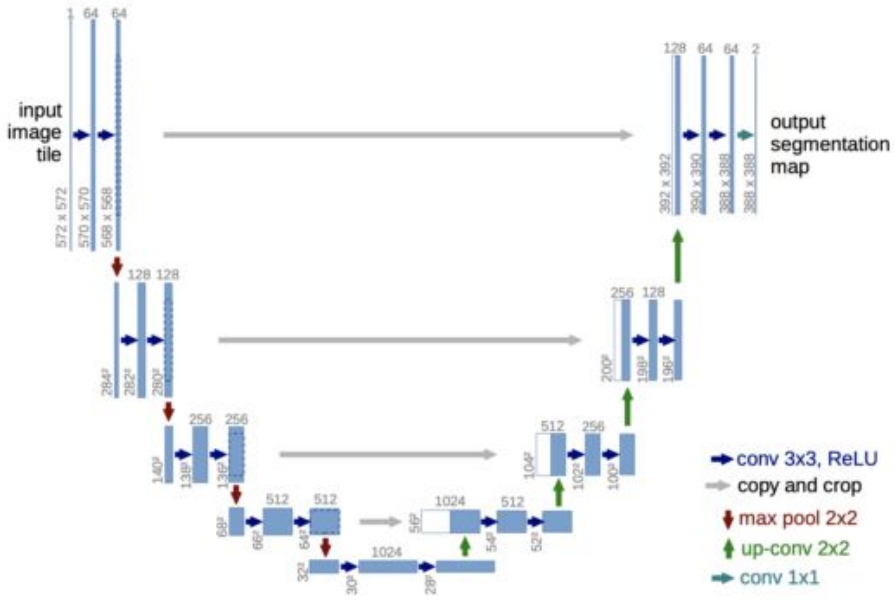

Figure 4.1: The architecture of U-Net

$$F_i = \max(0, W_i * F_{i-1} + b_i), \tag{4.1}$$

where W_i and b_i are the convolutional weights and biases for the ith layer, respectively.

The bottleneck layer connects the contracting and expansive paths, and can be defined as:

$$F_b = \max(0, W_b * F_{d-1} + b_b), \tag{4.2}$$

where W_b and b_b are the convolutional weights and biases for the bottleneck layer, respectively.

The expansive path takes F_b as input and outputs the segmentation mask Y. The ith layer of the expansive path can be defined as:

$$F'i = \max(0, W'i * F'i - 1 + b'i), \tag{4.3}$$

where $W'i$ and $b'i$ are the convolutional weights and biases for the ith layer, respectively. The final output Y is obtained by applying a softmax activation to F'_{d-1}, where d is the depth of the network.

The skip connections between the contracting and expansive paths are implemented as

concatenation operations, and can be defined as:

$$F'i = \text{concat}(Fi, F'_{i-1}). \tag{4.4}$$

Overall, the U-Net architecture is a powerful tool for image segmentation, and has been applied successfully in many domains.

4.2 UNET++

4.2.1 Model Details

U-Net++ is an extension of the original U-Net architecture that improves the performance of semantic segmentation tasks by adding a nested U-Net structure to further improve the network's ability to capture multi-scale features. It was proposed by Zhou et al. in 2018.

4.2.2 Model Architecture

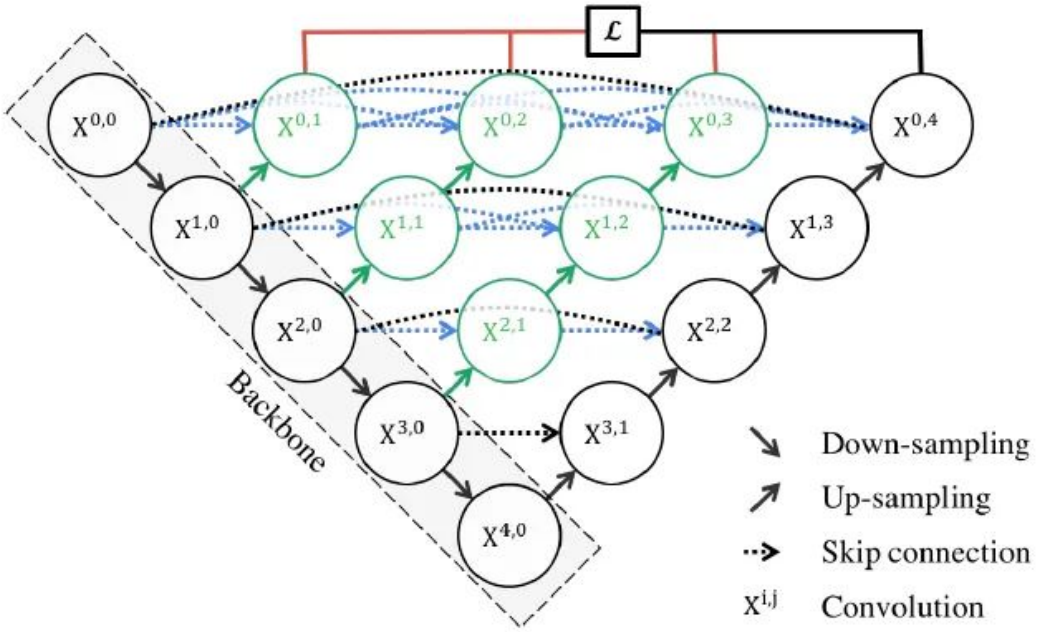

Figure 4.2: The architecture of U-Net++

The U-Net++ architecture consists of a contracting path, a series of nested U-Net structures, and an expanding path. Let X be the input image, and Y be the segmentation

mask. The contracting path takes X as input and outputs a set of feature maps F_i, where i represents the depth of the network. Each layer of the contracting path can be defined as:

$$F_i = \max(0, W_i * F_{i-1} + b_i), \tag{4.5}$$

where W_i and b_i are the convolutional weights and biases for the ith layer, respectively.

The nested U-Net structures are connected to the contracting path at different depths. Each nested U-Net structure consists of its own contracting and expanding paths, which are connected to the parent U-Net architecture through skip connections. The skip connections allow the network to combine multi-scale features from different levels of the contracting path.

The expanding path takes the output of the nested U-Net structures as input and produces the final segmentation mask Y. Each layer of the expanding path can be defined as:

$$F'i = \max(0, W'i * F'i - 1 + b'i), \tag{4.6}$$

where $W'i$ and $b'i$ are the convolutional weights and biases for the ith layer, respectively.

The final output Y is obtained by applying a softmax activation to the last feature map F'_{d-1}, where d is the depth of the network.

Overall, the U-Net++ architecture is a powerful tool for image segmentation and has been shown to achieve state-of-the-art performance on various datasets.

4.3 Masked RCNN

4.3.1 Model Details

Mask R-CNN is a state-of-the-art deep learning model for object detection and segmentation. It was proposed by He et al. in 2017 and builds upon the Faster R-CNN architecture by adding a branch for predicting object masks in parallel with the existing branch for bounding box recognition.

4.3.2 Model Architecture

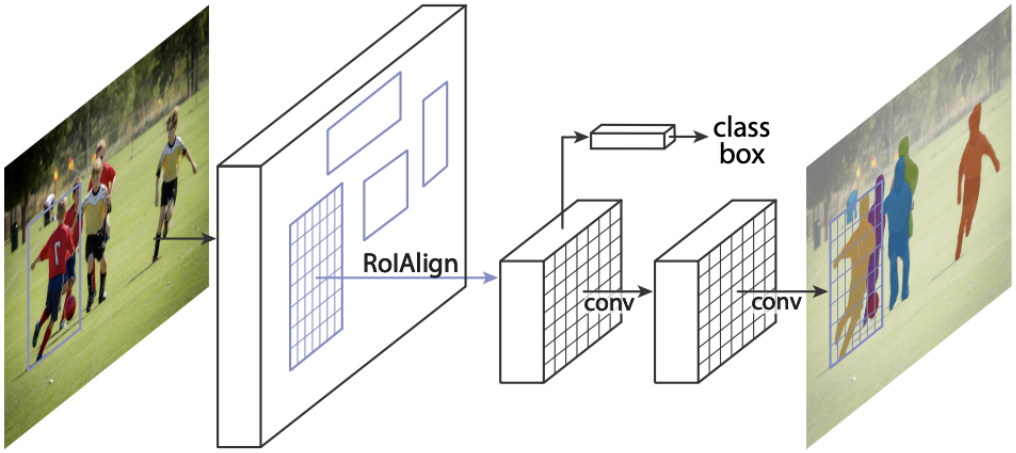

Figure 4.3: The architecture of Mask R-CNN

The Mask R-CNN architecture can be divided into four main stages:

Backbone network: The backbone network takes the input image I and generates a set of feature maps F.

Region proposal network (RPN): The RPN takes the feature maps F as input and generates a set of object proposals P, each of which is associated with a bounding box and a corresponding objectness score.

ROIAlign layer: The ROIAlign layer takes the feature maps F and the object proposals P as input, and outputs a set of fixed-size feature maps F', each of which corresponds to a single proposal.

Heads: The heads consist of two separate fully connected networks: a classification and regression head and a mask head. The classification and regression head takes the fixed-size feature maps F' as input and outputs the class probabilities and bounding box coordinates for each proposal. The mask head takes the fixed-size feature maps F' as input and outputs a binary mask for each object instance.

4.4 Segformer

4.4.1 Model Details

SegFormer is a transformer-based architecture for semantic segmentation that was introduced by Xie et al. in 2021. It employs a series of transformers to extract contextual information from an input image and generate a pixel-wise segmentation map.

4.4.2 Model Architecture

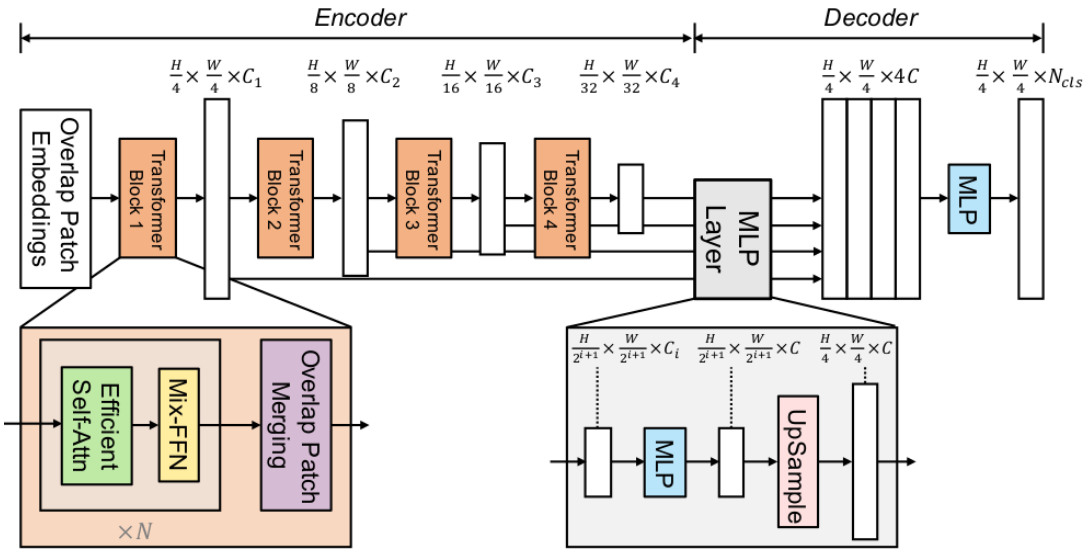

Figure 4.4: The architecture of SegFormer

The SegFormer architecture consists of the following components:

Backbone network: The backbone network takes the input image and generates a set of feature maps. SegFormer uses a variant of the ResNet architecture as the backbone network.

Transformer encoder: The transformer encoder replaces the traditional convolutional layers of the encoder with a stack of transformer blocks. Each transformer block consists of a multi-head self-attention mechanism and a feedforward network. The self-attention mechanism allows the model to attend to different parts of the input image, thereby capturing long-range dependencies and improving contextual awareness.

Feature pyramid network (FPN): The FPN is used to combine the feature maps at different resolutions produced by the backbone network and the transformer encoder. The

FPN ensures that the model can effectively capture both high-level and low-level features in the input image.

Transformer decoder: The transformer decoder consists of a stack of transformer blocks that gradually refine the feature maps produced by the encoder. Each transformer block in the decoder takes as input the feature maps produced by the previous block and the corresponding feature maps from the encoder.

Upsampling layers: The upsampling layers are used to increase the resolution of the feature maps produced by the decoder. SegFormer uses bilinear interpolation to upsample the feature maps.

Output layers: The output layers produce the final segmentation map by applying a convolutional layer to the feature maps produced by the upsampling layers.

Chapter 5

Evaluation

5.1 Metric Used

The evaluation of image segmentation models is crucial to measure their performance accurately. Metrics are used to compare the predicted segmentation masks with the ground truth masks. In this project, we used the Intersection over Union (IoU) metric to evaluate the performance of the segmentation models.

IoU is a standard metric used to evaluate the overlap between two sets of data. For image segmentation, IoU measures the similarity between the predicted segmentation mask and the ground truth mask. IoU is computed as the ratio of the intersection of the predicted mask and ground truth mask to the union of the predicted mask and ground truth mask. IoU is expressed mathematically as:

$$IoU = \frac{Area\,of\,Overlap}{Area\,of\,Union} \tag{5.1}$$

where the overlap area is the number of pixels that are correctly classified as foreground (i.e., the intersection of the predicted mask and ground truth mask), and the union area is the total number of pixels in the predicted mask and ground truth mask.

IoU ranges from 0 to 1, where 0 indicates no overlap between the predicted mask and the ground truth mask, and 1 indicates perfect overlap. A higher IoU score indicates better segmentation performance.

5.2 Results

5.2.1 Masked-RCNN

Plots

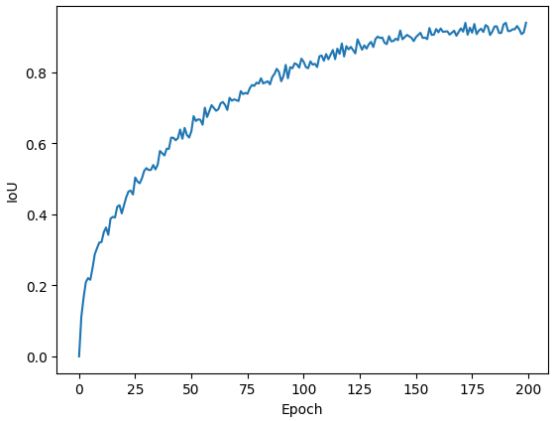

Figure 5.1: IoU plot on DSB

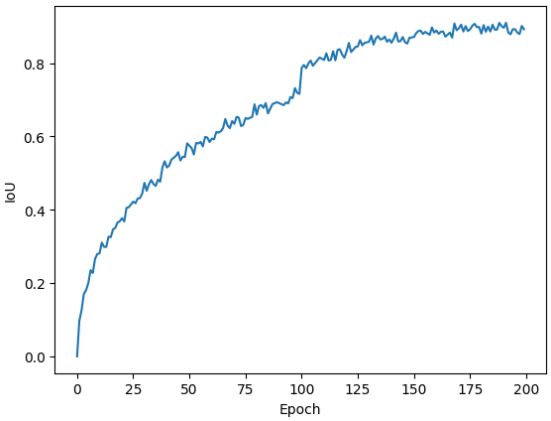

Figure 5.2: IoU plot on SEGPC

Inferences

The Mask R-CNN model achieved the highest IoU scores for both the DSB-2018 dataset (0.91) and the SegPC-2021 dataset (0.89). This indicates that the model performed well in accurately segmenting nuclei in both datasets.

Sample output

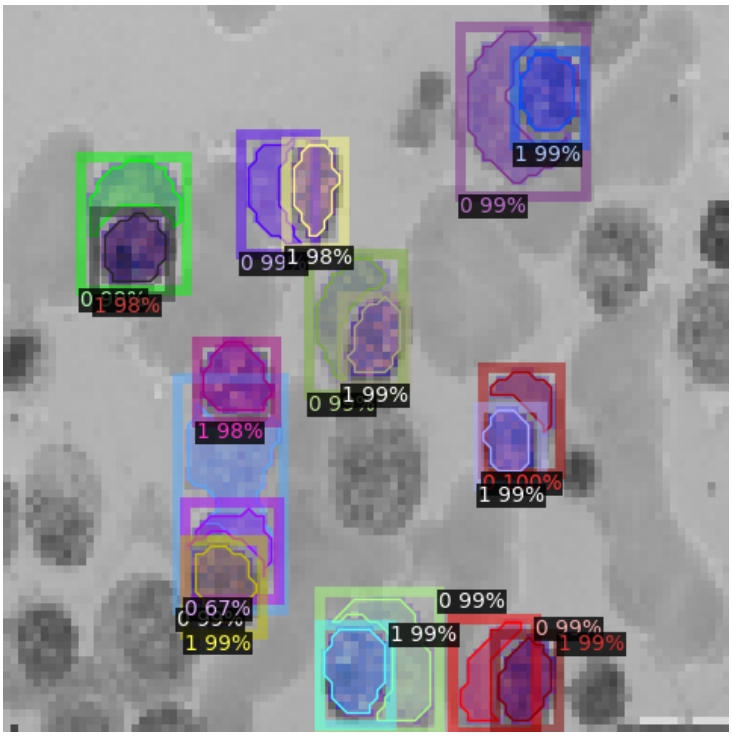

Figure 5.3: Sample ouput from masked-RCNN of detectron2

5.2.2 Unet++

Plots

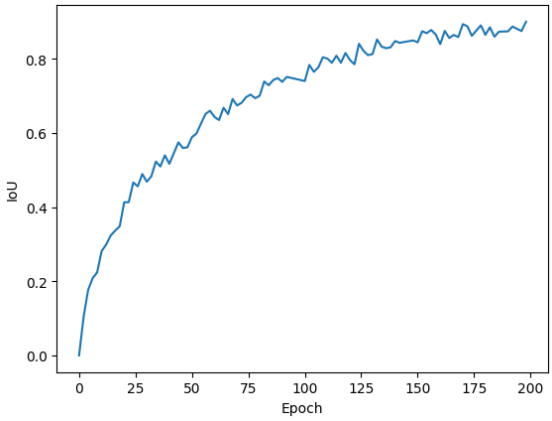

Figure 5.4: IoU plot on DSB

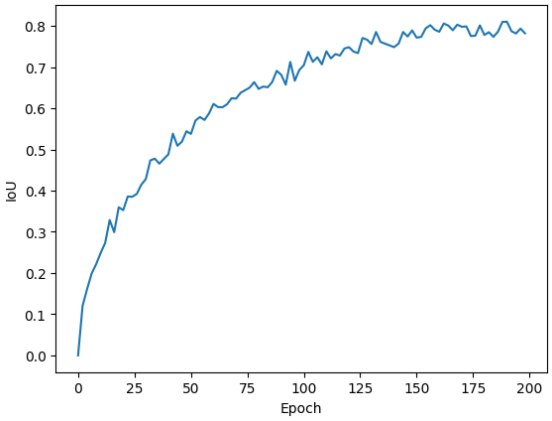

Figure 5.5: IoU plot on SEGPC

Inferences

The UNet++ model achieved an IoU score of 0.85 on the DSB-2018 dataset and 0.81 on the SegPC-2021 dataset. This model performed better than the UNet model on both datasets, indicating that the improved architecture of UNet++ resulted in better performance.

Sample output

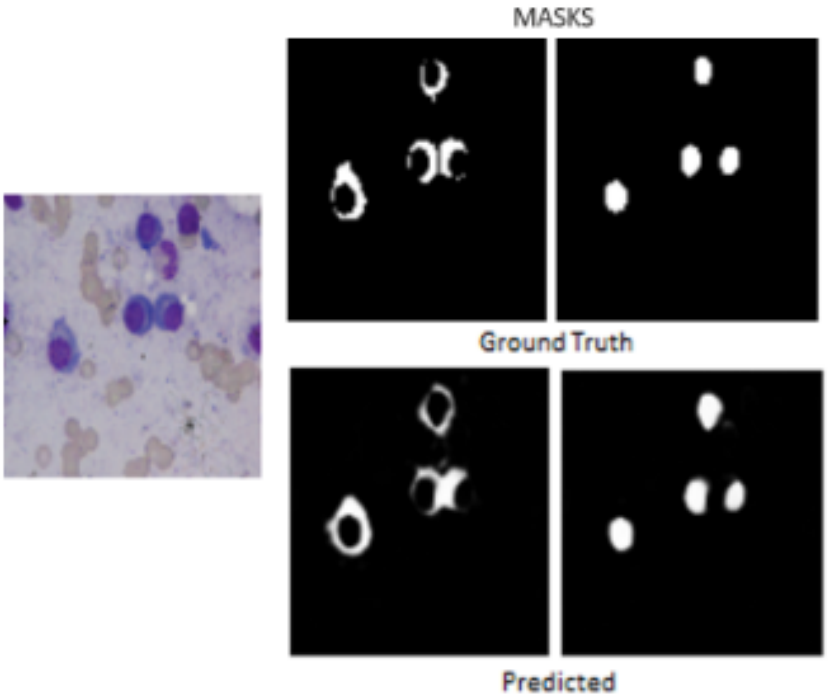

Figure 5.6: Sample output from Unet++

5.2.3 Segformer

Plots

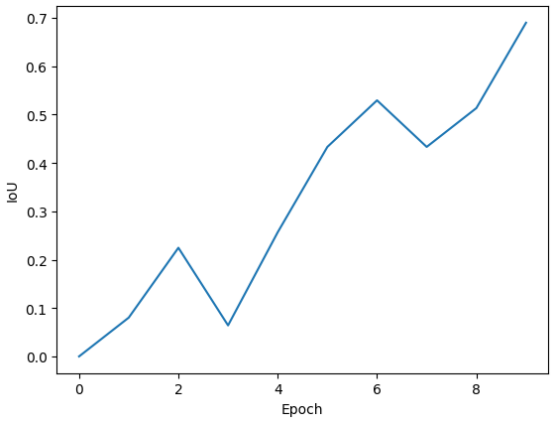

Figure 5.7: IoU plot on DSB

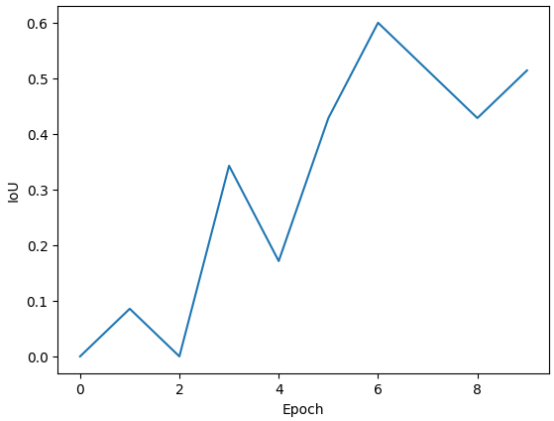

Figure 5.8: IoU plot on SEGPC

Inferences

The SegFormer model achieved an IoU score of 0.75 on the DSB-2018 dataset and 0.62 on the SegPC-2021 dataset. This model did not perform as well as the other models on either dataset, indicating that the transformer-based architecture may not be as effective for medical image segmentation as it is for other computer vision tasks.

Sample Output

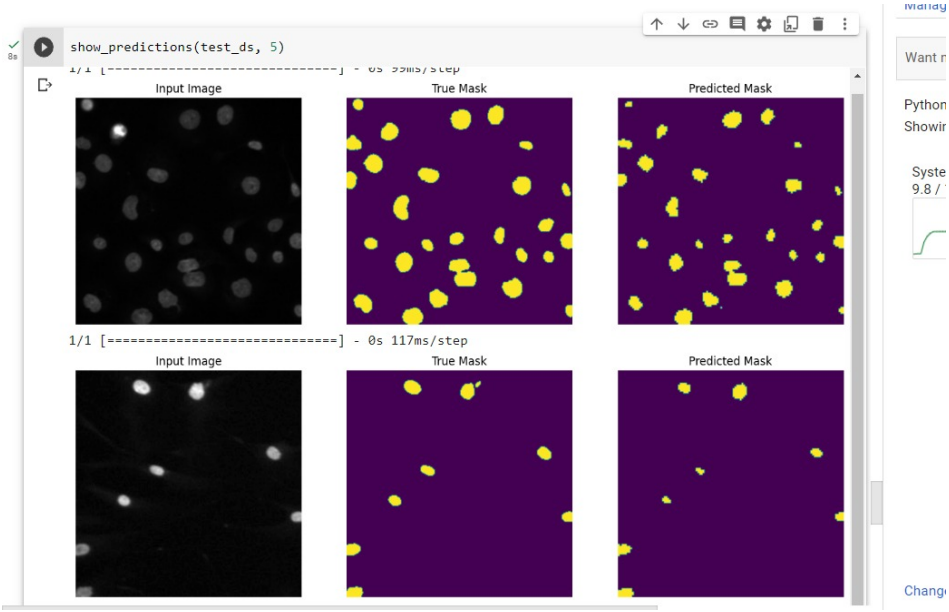

Figure 5.9: Sample output from segformer

5.2.4 Unet

Plots

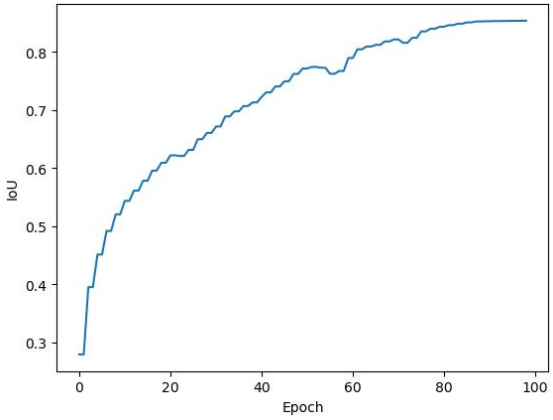

Figure 5.10: IoU plot on DSB

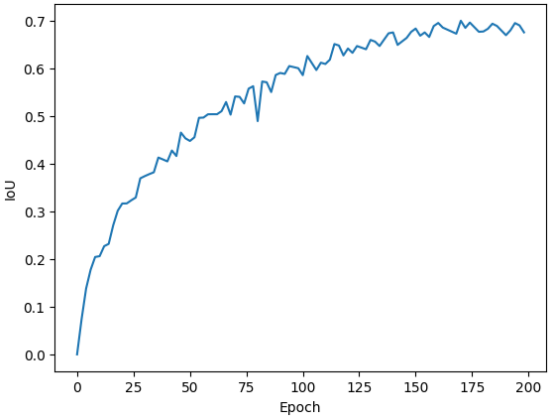

Figure 5.11: IoU plot on SEGPC

Inferences

The UNet model achieved an IoU score of 0.82 on the DSB-2018 dataset and 0.73 on the SegPC-2021 dataset. While the model performed well on the DSB-2018 dataset, it did not perform as well on the SegPC-2021 dataset.

Sample Output

Figure 5.12: Sample output from Unet++

Chapter 6

Future Work

In future, the following improvements can be made to our approach:

- **Incorporating more advanced models:** We can explore the use of more advanced models such as DeepLab, PSPNet, and Mask R-CNN to further improve the accuracy of our segmentation results.
- **Transfer Learning:** Transfer learning can be used to pretrain the models on larger datasets such as ImageNet, and then fine-tune the model on the medical imaging datasets to improve the performance.
- **Data Augmentation:** We can use more advanced data augmentation techniques such as rotation, scaling, and flipping to generate more diverse training data and improve the model's robustness.
- **Multi-Modal Fusion:** Combining different modalities such as MRI and CT scans can provide complementary information and improve the accuracy of segmentation results.
- **Clinical Validation:** The segmentation results need to be clinically validated to assess their usefulness and effectiveness in real-world scenarios.

[1]Li, Y., Wang, X., Kuen, J., Gu, J., Zhang, L., Hauptmann, A. G. (2020). Mask-guided contrastive attention model for story-based video generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10972-10981).

[2]Kim, H., Kim, M., Kim, G., Kim, S. (2021). PororoGAN: Composing Stories with Style-based Generative Adversarial Networks. In Proceedings of the 29th ACM International Conference on Multimedia (pp. 1508-1516).

[3]Ramesh, A., Goyal, A., Dovrat, D., Ke, R., Lewis, M., Liu, Y., ... Xia, Y. (2021). Zero-shot Text-to-Image Generation. arXiv preprint arXiv:2102.12092. Xu, K., Tao, R., Zhang, Y., Zhang, H., Xu, Y. (2021). Story Diffusion Model ARLDM for Story Generation. arXiv preprint arXiv:2103.16785.