Deep Learning Architectures for Medical Segmentation Tasks

Srinivas Natarajan
snatar28@asu.edu

***Abstract*— The importance of early detection localization of polyps (precursors to colon cancer) cannot be understated. The error-prone nature of the manual screening process for such abnormalities necessitates an automated system that can help pinpoint the location of anomalies with ease. In this project, we aim to exploit deep learning models that specifically employ Convolutional Neural Networks (CNNs) and their variants for their state-of-the-art performance in image identification tasks. The objective is to build a model that can be generalized to detect similar anomalies in such medical procedures and tag them reliably**

Keywords— Encoders, CNN, Segmentation, Pyramid Networks

I. Introduction

The goal of my project is to aid doctors in the identification of tumorous growths in the gastro-intestinal (GI) tract called polyps as they are precursors to colorectal cancer. This is an important problem in the medical field as colorectal cancer has the third highest mortality rate among cancers making it vital to identify early. Compounding this problem is the difficulty in identifying these growths as they are miniscule in size, blending in with the rest of the GI tract. It would be effective to have a computer-aided system to identify polyps and lesions during a medical procedure, especially when used in conjunction with a medical expert.

The general approach taken in medical segmentation problems is to extract key information from the original image by using Convolutional Neural Networks. This is due to their efficiency in processing larger images by extracting smaller features from them. Studies have found architectures such as UNet and ResUNet have shown great promise in tackling this problem but still require some tuning due to the contextual nature of the problem. However, CNN-based models in general show limitations for explicit long-range relations and they might exhibit unstable performance, unlike transformers. Therefore, we also use systems that employ encoder networks to extract information and decoder networks to interpret them into the form we need. The most successful approach of this form uses a series of contraction and expansion layers to downsize the initial image while preserving information and scaling it up.

II. Solution Explained

I first identified two major datasets collected by universities in conjunction with hospitals. They are: The Clinic CVC dataset [1] from the Endoscopy Vision challenge and the Kvasir-SEG dataset [2] collected by the Vestre Viken Health Trust (Norway), each consists of a sequence of still images taken from different endoscopic procedures. As this data is varied in terms of clarity, size and rotation, I first made a preprocessing pipeline to download these images, scale them to a 288x384 resolution to have an even padding with kernels of size 32 by either downscaling or padding them and normalized the pixel values to a 0-1 range to make computation easier. This processed data is then split into an 80-10-10 split for training, validation and testing respectively. Initial runs suffered from underfitting due to the lack in training data so I added an augmentation step into the pipeline. This involved flipping the training images horizontally and vertically, adding random rotations and crops and randomly altering the brightness. This also gave an added advantage of making the model more adaptable to real world conditions as there are many variables in this procedure.

With the pre-processing done, I then set up five different architectures with varying approaches to validate their feasibility. The first model was the UNet [3], a baseline standard for most medical image segmentation tasks. This symmetric architecture consists of two main components-contraction and expansion. The contraction section is a classic CNN architecture consisting of two 3x3 convolution layers, a ReLU and 2x2 max pooling unit. At each stride of the max pooling operation, we double the number of feature channels so in the expansion section, we have 2x2 up-convolution layer that halves the number of feature channels. Then, we have concatenation, two 3x3 convolution layers. Each of these layers are followed by a ReLU. Finally, we have 1x1 convolution layer to map each feature vector to a suitable number of classes.

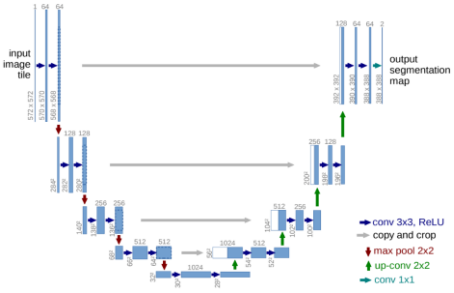

Figure: UNet Architecture

The next model implemented is the UNet++ [4] which adds residual gateways and skip pathways between the encoder and decoder to the UNet architecture. It employs deep supervision to perform more accurate segmentation. The re-designed skip pathways aim at reducing the semantic gap between the feature maps of the encoder and decoder sub-networks, and have proved effective in recovering fine-grained details of the target objects. The UNet model serves as a baseline as it the most common basepoint taken when tackling medical segmentation tasks. The UNet++ is taken as our first modification as it is a small upgrade over the base UNet but produces a much-improved results.

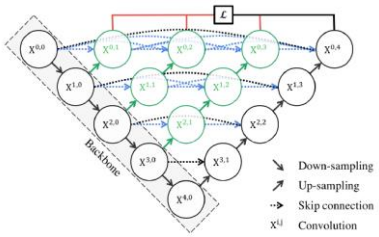

Figure: UNet++ Architecture

The next two models follow the methodology of pyramidal architectures of the encoder and decoder networks for segmentation. These are the Pyramid Attention Network (PAN) [5] and the Feature Pyramid Network (FPN) [6]. The Feature Pyramid Network produces proportionally scaled feature maps at each level. This map produced during the contraction phase is saved and used in the reconstruction of the more information dense image created during the expansion phase. It focuses on upscaling semantically stronger feature maps from higher levels to give the illusion of great feature resolution. It consists of two pathways: the bottom-up pathway usually made up of CNNs to extract features where the spatial resolution decreases but semantic values increases and the top-down pathways where the object is down sampled. There are lateral connections between the reconstruction layers and the feature maps to better predict locations.

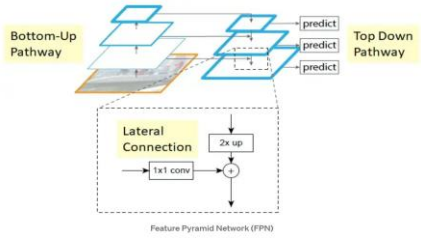

Figure: Feature Pyramid Network Architecture

The Pyramid Attention Network uses two important modules: the Global Attention Upscale (GAU) and the Feature Pyramid Attention (FPN). The FPN modules is a U-shaped architecture that extracts information using 3x3, 5x5, 7x7 filters. The pyramid structure uses information extracted every step without the additional burden of using larger filters as the resolution of the feature map is already small. The GAU unit performs global average pooling to provide global context to the process. This is then passes through a 1x1 convolution filter, batch normalized and passed through a ReLU activation function to add non-linearity.

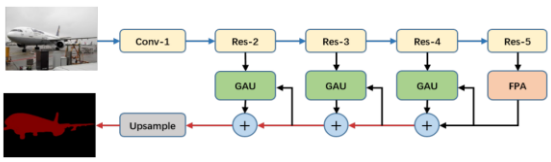

Figure: Pyramid Attention Network Architecture

With the pyramid extraction methods out of the way, the last model left is one that represents the current state of the art work. For this we chose the DeepLabv3 architecture [7] developed by Google. This novel approach solves the information loss when in two ways. When downscaling the image, a combination of atrous convolutional network and spatial pyramid pooling modules helps reduce the info loss. The advantage of this method is that through the dilation parameter, the network can extract the information that could be potentially obtained with a larger convolution kernel while avoiding the cost that comes with actually using a larger kernel. While upscaling, the use of a technique called Point Rend enhancement helps reduce the loss of info in the process.

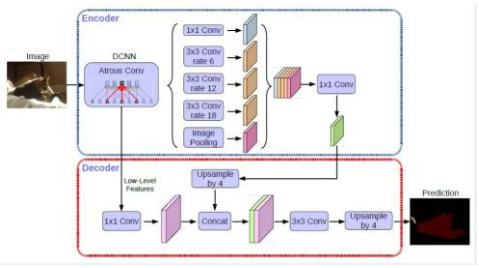

Figure: DeepLabv3 Architecture

III. Results

The five model were used to predict polyp boundaries on our test dataset. We used two metrics to compare these models: IoU and Dice scores. The IoU score which stands for Intersection over Union, signifying the ratio of the intersecting region over the overlapping region. An IoU score of 1 indicates a perfect boundary prediction. The IoU score can be given as follows:

$$IoU = \frac{TP}{TP + FP + FN}$$

Equation: Intersection over Union Score

It calculates the sum of correctly predicted boundary pixels over the sum of total boundary pixels of both prediction and ground truth. This is a good metric to use as it considers both local and global information loss. The Dice loss can be given as follows:

$$D = \frac{2\sum_i^N p_i * g_i}{\sum_i^N p_i^2 + \sum_i^N g_i^2}$$

Equation: Dice Loss

We found that the best models among the ones tested were the UNet++ and the DeepLabv3.

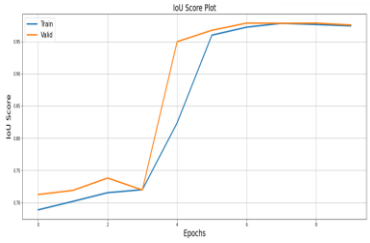

Figure: UNet++ IoU score

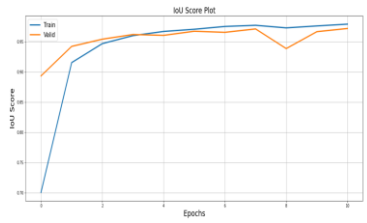

Figure: DeepLabv3 IoU score

Though the results were close when we look at the IoU scores, coming withing a margin of error. We can see that the DeepLabv3 model with the help of a better upscaling algorithm is able to make better use of the data available by better maintaining the location details in the image. This is clear in the time it takes for the IoU score to plateau, especially compared to the UNet++ score in the first few epochs. Let's compare the dice losses for a more definitive.

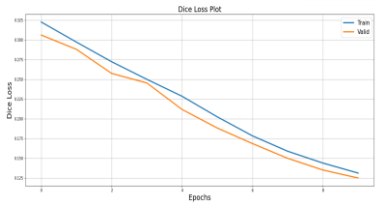

Figure: UNet++ Dice Loss

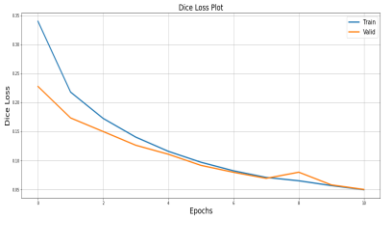

Figure: DeepLabv3 Dice Loss

We use the Dice score as a tie breaker as it penalizes mistakes more heavily. This is critical when developing any application for medical purposes as the margin of error is minimal. With this we come to the conclusion that the best model to use is the DeepLabv3.

Table I
Performance Comparisons of Models

Model	IoU Score	Dice Loss
UNet	0.9684	0.0475
UNet++	0.9756	0.1250
FPN	0.8920	0.0574
PAN	0.9526	0.0326
DeepLab v3	0.9716	0.0500

Figure: Performance comparisons of the model

In conclusion, we explained the shortcoming of each architecture as well as specific use cases where they shine. We found that the DeepLabv3 architecture was the best in terms of our metrics which emphasized minimizing the error rate. We went forward with this metric as the margins for error in the medical field are miniscule and errors should be heavily penalized.

IV. Contribution and Skills

My role in this project included:

1. Creating the data collection, preprocessing and augmentation pipeline for two datasets of different formats.
2. Implementing the DeepLabv3 architecture from the research paper recently published by Google using the PyTorch library and its functions.
3. Designing the graphing and results collection function for the various models to come to a cohesive result.

The other members involved in this group project are Sai Pranav Tavva (1225344341), Tanushi Ahuja (1225475680), Pranavi Addagatla (1225696667) and Shivani Yerram (1225766373)

Over the course of this project, I learnt more about processing images for training deep learning models. This was vital as the amount of processing power needed to effectively tackle these problems is immense and the data needs to be modified to make it feasible in finite time and ordinary hardware. I also learned about the various encoder-decoder architectures for image feature extraction and their strengths and weaknesses for various applications. Through experimentation with the PyTorch package when creating the various model architectures, I also learned about additional packages which work with PyTorch like the "Albumnations" and the "Segmentation" packages which offer important functionalities which made creating data pipelines much easier.

V. References

[1] Bernal, J., Sánchez, F. J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., & Vilariño, F. (2015). WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Computerized Medical Imaging and Graphics, 43, 99-111.

[2] 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," MultiMedia Modeling, pp. 451–462, 2019.

[3] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," arXiv.org, 18-May-2015. [Online]. Available: https://doi.org/10.48550/arXiv.1505.04597.

[4] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh and J. Liang, "UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation," in IEEE Transactions on Medical Imaging, vol. 39, no. 6, pp. 1856-1867, June 2020, doi: 10.1109/TMI.2019.2959609.

[5] H. Li, P. Xiong, J. An, and L. Wang, "Pyramid Attention Network for semantic segmentation," arXiv.org,(2018).https://doi.org/10.48550/arXiv.1805.10180.

[6] Taresh Sarvesh Sharan, Sumit Tripathi, Shiru Sharma & Neeraj Sharma (2022) Encoder Modified U-Net and Feature Pyramid Network for Multi-class Segmentation of Cardiac Magnetic Resonance Images, IETE

Technical Review, 39:5, 1092-1104, DOI: 10.1080/02564602.2021.1955760
[7] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for Semantic Image segmentation,” Computer Vision – ECCV 2018, pp. 833–851, 2018.