Polyp Segmentation Project Report

Problem Statement

The primary objective of this project is to develop a machine learning model for the segmentation of gastrointestinal polyps in medical images. The underlying question is: "Can we automate the identification and delineation of polyps in gastrointestinal images to assist healthcare professionals in early diagnosis and treatment planning?" The project adds significant societal value by potentially reducing the time and effort required for manual segmentation, thereby expediting the diagnostic process and improving patient outcomes.

Background on the Subject Matter

The segmentation of gastrointestinal polyps is a critical step in the early detection of conditions that could lead to colorectal cancer. Traditional methods involve manual inspection of endoscopic images by healthcare professionals, which is time-consuming and subject to human error. The application of data science techniques, particularly machine learning and computer vision, can automate this process, making it faster and more accurate. In the past, various algorithms and techniques like thresholding and edge detection were used, but they lacked the sophistication and accuracy that modern machine learning models can provide.

Details on Dataset

Source and Structure

The dataset used for this project is the Kvasir-SEG dataset, a publicly available collection of gastrointestinal polyp images and corresponding masks hosted by Simula Datasets. The dataset includes images and their corresponding segmentation masks JPEG format. Additionally, the dataset provides bounding box information for each image, stored in a JSON file.

Collection

The dataset was collected by medical professionals during colonoscopy procedures and is intended for research in the field of computer-aided diagnosis. It aims to assist in the early detection and treatment of gastrointestinal polyps, which can be precursors to cancer.

Schema

- **Images**: Stored in JPEG format, each image represents a unique instance of a gastrointestinal polyp.
- **Segmentation Masks**: Stored in JPEG format, these grayscale images serve as the ground truth for segmentation.
- **Bounding Boxes**: The coordinate points for the bounding boxes around the polyps are stored in a JSON file. Each entry in the JSON file corresponds to a specific image and contains the coordinates that define the bounding box.

Summary of Cleaning and Preprocessing

Exploratory Data Analysis (EDA)

Before diving into model building, a comprehensive EDA was conducted to understand the distribution of polyp sizes, their locations within the images, and the color distributions. This helped in fine-tuning the data augmentation strategies and provided insights into the complexities of the problem at hand.

Preprocessing

1. **Rescaling**: The pixel values of the images were rescaled to a range of 0 to 1 by dividing each pixel by 255. This normalization helps the model converge faster during training.
2. **Resizing**: All images and masks were resized to a uniform dimension of 256x256 pixels. This standardization ensures that the model receives inputs of consistent size, which is crucial for the convolutional layers to function correctly.

Data Augmentation

To improve the model's ability to generalize, various data augmentation techniques were applied to the training set. These include horizontal and vertical flips, rotation, zooming, and brightness adjustments. The motivation behind using data augmentation is to artificially increase the diversity of the training data, thereby making the model robust to various transformations.

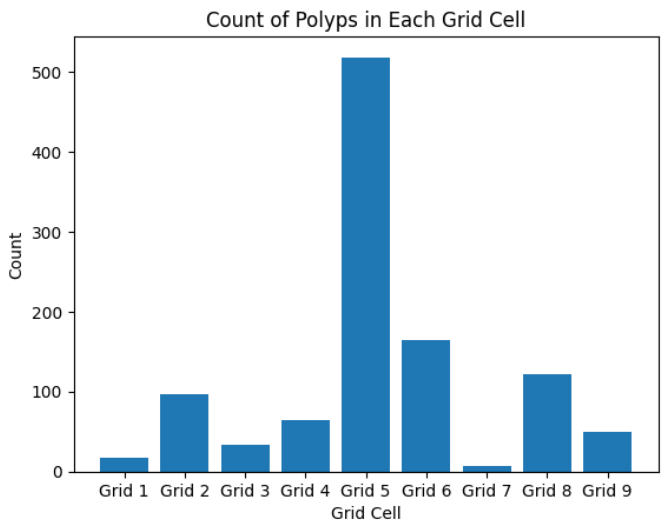
Insights, Modeling, and Results

Insights from EDA

Polyp Position

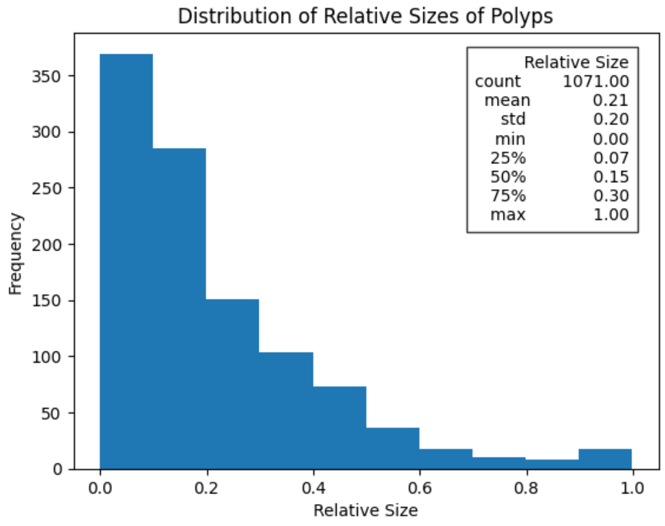
To understand the spatial distribution of polyps within the images, each image gets divided into a 3x3 grid. Iterating over the bounding box data allows for the calculation of the center coordinates of each bounding box, determining its corresponding grid cell. The analysis reveals a predominant occurrence of polyps in the center of the images, specifically in the top-center (Grid2), center-center (Grid5), and bottom-center (Grid8) regions. Conversely, the left side of the images—comprising the top-left (Grid1), center-left (Grid4), and bottom-left (Grid7) regions—shows the fewest occurrences of polyps. This insight proves valuable for targeted image analysis and influences the design and evaluation of the segmentation model.

Grid 1	Grid 2	Grid 3
Grid 4	Grid 5	Grid 6
Grid 7	Grid 8	Grid 9



Polyp Size

To quantify the relative size of the polyps, the area of each bounding box is divided by the total area of the corresponding image. The area of a bounding box is calculated by multiplying its width and height. The analysis shows that the majority of the polyps occupy between 7% to 30% of the total image area. Instances where polyps occupy more than 50% of the image are rare. Interestingly, the dataset contains 1,071 polyps across 1,000 images, indicating that some images feature multiple polyps. This observation is crucial for understanding the complexity of the segmentation task and informs the model's design and evaluation criteria.



	Relative Size
count	1071.00
mean	0.21
std	0.20
min	0.00
25%	0.07
50%	0.15
75%	0.30
max	1.00

Color Distribution

The color distribution of the polyps was analyzed using the HSV (Hue, Saturation, Value) color space, and the following observations were made:

Hue:

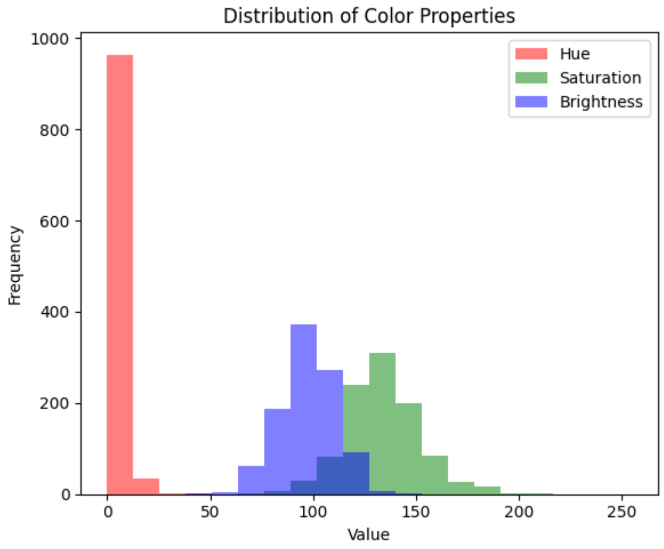
The hue values range from 4.28 to 48.53, with most values falling within the 25th percentile of 7.82 and the 75th percentile of 9.71. This range predominantly corresponds to the red color spectrum. A longer tail towards higher hue values indicates a lack of color bias in that direction.

Saturation:

Saturation values span a wide range from 75.41 to 206.39, with a mean value of 134.14. The distribution appears fairly symmetric, suggesting a moderate level of saturation on average across the dataset.

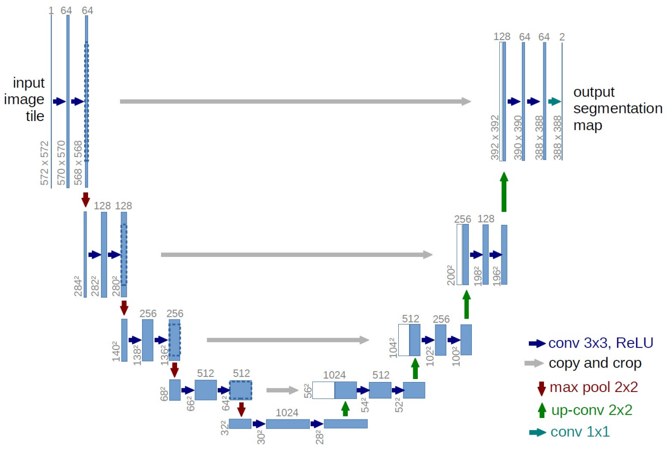
Brightness:

Brightness values range from 48.13 to 151.25, with a mean value of 97.48. The distribution is fairly symmetric, indicating no strong bias towards either high or low brightness levels.



Model Selection

Given the complexity and the need for high accuracy in medical image segmentation, the U-Net architecture was chosen for this project. U-Net is renowned for its effectiveness in biomedical image segmentation tasks, offering a good balance between computational efficiency and performance.



Model Evaluation Metrics

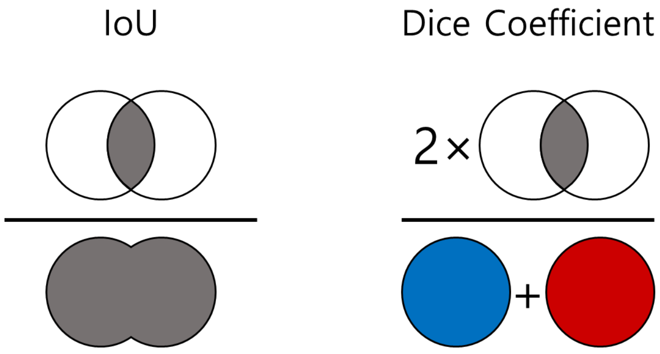
The model was evaluated using multiple metrics, including accuracy, precision, recall, F1-score, IoU (Intersection over Union), and Dice Coefficient. These metrics provided a comprehensive understanding of the model's performance, especially in terms of its ability to correctly identify and delineate polyps.

```
              precision    recall  f1-score   support

  Background       0.96      0.98      0.97   5451819
  Foreground       0.91      0.80      0.85   1101781

    accuracy                           0.95   6553600
   macro avg       0.93      0.89      0.91   6553600
weighted avg       0.95      0.95      0.95   6553600
```

- **Accuracy**: 95%
- **Precision**: 91% for Foreground (Polyps)
- **Recall**: 80% for Foreground (Polyps)



- **IoU**: 0.738
- **Dice Coefficient**: 0.857

Results

The model demonstrated high accuracy and precision, indicating its reliability in identifying polyps. However, the recall was slightly lower, suggesting room for improvement in capturing all potential polyps, especially the smaller ones. The IoU and Dice Coefficient also indicated a strong overlap between the predicted and ground-truth masks, affirming the model's effectiveness.

Findings and Conclusions

The project aimed to develop a robust polyp segmentation model to assist healthcare professionals in early diagnosis and treatment planning. The initial hypothesis was that leveraging advanced machine learning techniques would yield a model with high accuracy, precision, and recall. The results largely align with these initial goals, achieving a high degree of morphological similarity to the ground truth in segmentation tasks.

However, the model does exhibit some limitations, such as the presence of spurious detections or false positives, affecting its precision. These findings suggest that while the model is effective, there is room for optimization, particularly in the areas of spatial accuracy and reducing false positives.

In terms of practical value, the project has significant implications for healthcare, especially in the early detection of potentially cancerous polyps. The technology can be integrated into existing medical imaging systems to provide real-time, accurate segmentation, thereby aiding in quicker and more effective treatment planning.

Practical Applications

- Integration into medical imaging systems for real-time polyp segmentation.
- Assisting radiologists and healthcare professionals in early and accurate diagnosis.

Future Directions

- Further optimization of the model to improve spatial accuracy and reduce false positives.
- Extending the model to handle other types of medical images and segmentation tasks.