GI Image Classification with ResNet-18

Brian Samir Tiscareño Bisteni
Universidad Panamericana
Aguascalientes, México
0252756@up.edu.mx

Introduction

The human body can be affected by several diseases in the digestive system. Currently endoscopic and gastroscopy examinations are made to identify illnesses; the endoscopy focuses on the gastrointestinal tract while gastroscopy aims for the stomach and first part of the small bowel. These processes are highly dependent on the doctor's insight, although doctors are qualified and capable of discerning dangerous tissues, errors could happen and may affect the patient's treatment. Therefore, automatic detection and recognition have been considered, these tools are helpful to determine an accurate grading of disease and ease the decision-making process.

With this innovation, medicine is going through a profound transformation. A well-trained model has the potential to outperform doctors in these certain tasks, significantly reducing both costs and time. Given that time is crucial in medical contexts, faster responses to health issues can lead to timely interventions, consequently improving patients' quality of life.

Dataset description

Data was collected from the Vestre Viken Health Trust (VV) in Norway, specifically from the Bærum Hospital, this hospital stands out for its large gastroenterology department, the images are carefully annotated by one or more medical experts from VV and the Cancer Registry of Norway (CRN).
The multiclass dataset consists of images with different sizes from 720x576 to 1920x1072 pixels that show anatomical landmarks, pathological findings or endoscopic procedures in the GI tract.
The dataset categorizes images into:
Three anatomical landmarks: Z-line (esophagus-stomach transition), pylorus (stomach-duodenum opening), and cecum (proximal large bowel with appendiceal orifice).
Three pathological findings: Esophagitis (esophageal inflammation), polyps (mucosal outgrowths with cancer potential), and ulcerative colitis (chronic inflammation of the large bowel).
Two categories related to polyp removal: Dyed and lifted polyps, and dyed resection margins, used in endoscopic mucosal resection (EMR).
These classes cover critical aspects of GI endoscopy, supporting both single- and multi-disease detection research.

Methodology

Initially the data set is divided into 70% training, 15% validation and 15% for testing, succeeding preprocessing was applied, consisting of the following procedures:

Resize: Training images were resized to 224 x random zoom between 0.8x to 1.2x, validation and testing images were resized to 224 X 224.

NumPy array to PyTorch tensor: Transforms the image into a format suitable for PyTorch, a tensor of shape [C, H, W], where C is channels, H is height, W is width, for example [3, 224, 224] for RGB using ToTensor() function.

Normalization: Image statistics were aligned wit the ones of ImageNet dataset, the same dataset with which ResNet-18 was pre-trained by default. The tensor values were normalized using the mean and standard deviation for each RGB channel.

Flipping images: Randomly flipped the images horizontally with a 50% probability, to increase variety and help the model generalize better.

Histogram equalization: Implemented a custom histogram equalization function using OpenCV, converted the images to YUV color space, equalized the Y (luminance) channel, and converted them back to RGB to enhance contrast.

Zoom: Randomly cropped and resized the images to 224x224 with a scale factor between 0.8x and 1.2x. with the intention to act as a zoom augmentation, simulating different magnifications and improving robustness.

The next step was to define and configure the model ResNet-18, the pretrained version from ImageNet, with Modify FC Layer replaced the final fully connected layer to output 8 classes for the Kvasir dataset.

ResNet-18 is a convolutional neural network to classify images, feature extraction and transfer learning. In this case it is applied to detect and analyze the GI diseases, it is useful for object detection or segmenting regions in images.

The convolutional layers extract hierarchical features like edges on low levels or polyp patterns in high levels. The residual blocks assure the features to accumulate without degradation; the output is reduced to 512 features via pooling. The new layer maps these classes to 8 punctuations without normalizing per feature (logits).

Loss is calculated by comparing the logits with the real labels, penalizing wrong predictions, ResNet-18 adjusts weights to maximize the correct class probability, SGD with a 0.9 momentum and a learning rate of 0.001 which updates weights based on the loss gradient. For very batch a “running_loss” variable is accumulated and divided by the total amount of images to obtain a global mean for the model’s performance

The model is trained for 4 epochs, an epoch is a complete iteration through the entire training process. In other words, an epoch means that each image in the training set has been used once to update the model's weights.

The model uses 8-size batches to divide the training dataset in groups of 8 images. For every batch, it receives the inputs and labels, calculates the predictions, computes the loss and updates the weights.

An epoch occurs when all batches in the training set have been processed once. During this process, the model adjusts its weights based on the gradients calculated for each batch, gradually refining its ability to predict the correct classes.

For every epoch the following metrics are calculated for the validation and testing partitions:

True positive (TP) The number of correctly identified samples.

True negative (TN) The number of correctly identified negative samples.

False positive (FP) The number of wrongly identified samples.

False negative (FN) The number of wrongly identified negative samples.

Recall (REC): Probability of detection and true positive rate, and it is the ratio of samples that are correctly identified as positive among all existing positive samples.

Precision (PREC): Shows the ratio of samples that are correctly identified as positive among the returned samples.

Specificity (SPEC): Shows the ratio of negatives that are correctly identified as such.

Accuracy (ACC): The percentage of correctly identified true and false samples.

Matthews correlation coefficient (MCC): It takes into account true and false positives and negatives, and is a balanced measure even if the classes are of very different sizes.

F1 score (F1): Measures a test's accuracy by calculating the harmonic mean of the precision and recall.

For validation and testing macro-averaged metrics are calculated as well.

The latest epoch is the one with lowest loss, meaning it is the one with the most accurate metrics obtained.

Experiments and results

The main idea was to employ a ResNet-50 architecture alongside a CUDA-enabled version of PyTorch, taking advantage of its capacity to manage a greater number of layers, its efficient bottleneck design, the extensive parameters it can utilize, and its superior accuracy. Incorporating CUDA would have significantly reduced computational time, notwithstanding the substantial resource demands imposed by ResNet-50.
During the development of the code to address these challenges, I encountered several limitations due to the modest capabilities of my computer. Initially, I planned to implement 10 epochs with a batch size of 32; however, my system's hardware and the lack of a GPU proved incapable of handling such a high number of iterations. Consequently, I adjusted the parameters to use only 3 epochs and a batch size of 8.
Despite the reduced number of iterations, the results revealed a notably slow training process, yet the model demonstrated robust performance.

4th epoch

Validation Metrics Table

Class	TP	TN	FP	FN	Precision	Recall	Specificity	Accuracy	F1 Score	MCC
dyed-lifted-polyps	101	1049	1	49	0.9902	0.6733	0.9990	0.9583	0.8016	0.7974
dyed-resection-margins	149	1006	44	1	0.7720	0.9933	0.9581	0.9625	0.8688	0.8565
esophagitis	124	1028	22	26	0.8493	0.8267	0.9790	0.9600	0.8378	0.8151
normal-cecum	146	1043	7	4	0.9542	0.9733	0.9933	0.9908	0.9637	0.9585
normal-pylorus	149	1048	2	1	0.9868	0.9933	0.9981	0.9975	0.9900	0.9886
normal-z-line	130	1025	25	20	0.8387	0.8667	0.9762	0.9625	0.8525	0.8311
polyps	139	1043	7	11	0.9521	0.9267	0.9933	0.9850	0.9392	0.9307
ulcerative-colitis	143	1039	11	7	0.9286	0.9533	0.9895	0.9850	0.9408	0.9323
Macro-Averaged	-	-	-	-	0.9090	0.9008	0.9090	0.9008	0.8993	0.8882

Test Metric table

Class	TP	TN	FP	FN	Precision	Recall	Specificity	Accuracy	F1 Score	MCC
dyed-lifted-polyps	113	1044	6	37	0.9496	0.7533	0.9943	0.9642	0.8401	0.8272
dyed-resection-margins	146	1014	36	4	0.8022	0.9733	0.9657	0.9667	0.8795	0.8658
esophagitis	110	1022	28	40	0.7971	0.7333	0.9733	0.9433	0.7639	0.7326
normal-cecum	149	1043	7	1	0.9551	0.9933	0.9933	0.9933	0.9739	0.9703
normal-pylorus	149	1050	0	1	1.0000	0.9933	1.0000	0.9992	0.9967	0.9962
normal-z-line	123	1009	41	27	0.7500	0.8200	0.9610	0.9433	0.7834	0.7519
polyps	137	1046	4	13	0.9716	0.9133	0.9962	0.9858	0.9416	0.9341
ulcerative-colitis	144	1043	7	6	0.9536	0.9600	0.9933	0.9892	0.9568	0.9506
Macro-Averaged	-	-	-	-	0.8974	0.8925	0.8974	0.8925	0.8920	0.8780

Test Confusion Matrix

Test Confusion Matrix

True \ Predicted	dyed-lifted-polyps	dyed-resection-margins	esophagitis	normal-cecum	normal-pylorus	normal-z-line	polyps	ulcerative-colitis
dyed-lifted-polyps	113	36	0	0	0	0	1	0
dyed-resection-margins	4	146	0	0	0	0	0	0
esophagitis	0	0	110	0	0	40	0	0
normal-cecum	0	0	0	149	0	0	1	0
normal-pylorus	0	0	0	0	149	1	0	0
normal-z-line	0	0	27	0	0	123	0	0
polyps	2	0	0	4	0	0	137	7
ulcerative-colitis	0	0	1	3	0	0	2	144

1st epoch -0.6422 loss
2nd epoch -0.3585 loss
3rd epoch - 0.2879 loss
4th epoch - 0.2316 loss

Conclusion

Despite the hardware limitations, which necessitated reducing the number of epochs from 10 to 4 and the batch size from 32 to 8, the model demonstrated commendable performance. The latest macro-averaged test accuracy at Epoch 4 reached 0.8925 (89.25%), surpassing the heuristic estimate of 0.7684 derived from 1 - 0.2316, the final training loss. This discrepancy highlights that while 1 - loss offers a rough approximation of model correctness, the true accuracy is more accurately reflected by the confusion matrix and metrics such as true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN). These metrics underscore the model's ability to generalize across the eight classes of the Kvasir dataset, despite the constraints, suggesting that further optimization with enhanced computational resources could yield even higher accuracy.

References

Kaggle. (n.d.). *Recommender system using Amazon reviews* [Computer code]. Retrieved June 4, 2025, from https://www.kaggle.com/code/saurav9786/recommender-system-using-amazon-reviews/input