DI725-Project-Second Phase

1st Emre Çakmakyurdu
Middle East Technical University
Data Informatics
Ankara, Turkey
emre.cakmakyurdu@metu.edu.tr

***Abstract*—**

***Index Terms*—**

I. Introduction

Ulcerative colitis (UC) is a chronic inflammatory disease affecting the colon, often diagnosed and monitored through endoscopic evaluations. The Mayo Endoscopic Scoring (MES) system is widely used to classify the severity of UC based on visual inspection, ranging from Mayo-0 (healthy) to Mayo-3 (most severe). Accurate classification is crucial, yet challenging, due to the inherent subjectivity in human interpretation and the presence of class imbalance in the available data.

This study aims to leverage transformer models to enhance the diagnostic process for UC. Given the ordinal nature of the MES levels and the imbalance between classes, the quadratic weighted kappa (QWK) metric is used as the primary evaluation measure. The proposed approach will strive to exceed baseline methods while considering the particularities of class distribution and ordinality. Additionally, the model's performance will be reported through confusion matrices for a comprehensive assessment.

The LIMUC dataset, consisting of endoscopic images, will serve as the basis for this classification task. All images are uniformly sized at 352x288 pixels, enabling consistency in input data processing. The ultimate goal is to develop a transformer-based model that outperforms existing baselines and offers valuable assistance to medical professionals in diagnosing and managing UC.

II. Dataset

LIMUC dataset is a publicly available Ulcerative Colitis dataset which is labeled according to Mayo Endoscopic Score (MES). MES has several labels (0-3), three represents the most severe case of the UC and zero represents the most minor case. There are 11276 352x288 pixel images from 564 patients and all images have been reviewed and annotated by experts. The distribution of classes are as follows:

- Mayo 0: 6105 (54.14%)
- Mayo 1: 3052 (27.70%)
- Mayo 2: 1254 (11.12%)
- Mayo 3: 865 (7.67%)

15% of the data is used as test set and the rest is used for train and validation set. Validation splitting size is set to be 0.1 which corresponds 8.5% of the total images are used as validation set whereas 76.5% of the images is used as training set. No k-fold cross validation is applied to the dataset. The images of the dataset for each class is as follows.

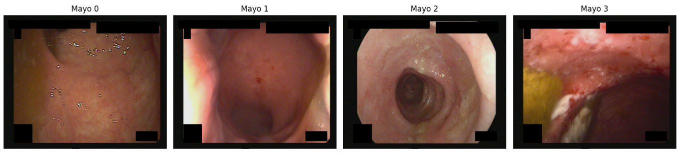

Fig. 1. Sample images of each Mayo index.

III. Literature Review

A. Class Distance Weighted Cross-Entropy Loss for Ulcerative Colitis Severity Estimation

This paper tackles the Ulcerative Colitis (UC) severity classification problem by introducing a new loss function called Class Distance Weighted Cross Entropy (CDW-CE) which can replace the cross entropy loss for ordinal classification tasks. Scoring system used in this task is Mayo Endoscopic Score (MES) labels starting from 0-3 where 3 represents the most severe case. CDW-CE is a non parametric loss function which penalizes incorrect predictions according to their distance from the correct class.

Paper uses LIMUC dataset which is a collection of UC images according to MES. Authors use 10-fold cross validation they make all the splittings at the patient level by preserving class ratios.

$$\text{CDW-CE} = -\sum_{i=0}^{N-1} \log(1-\hat{y}_i) \times |i-c|^{\alpha} \tag{1}$$

Authors trained three different models with different loss functions such as CDW-CE, Cross-Entropy, cross entropy with an ordinal loss term, ordinal entropy loss and CORN framework. Trained models are ResNet18, Inception-v3 and MobileNet-v3-large respectively. Authors applied preprocessing steps such as random rotation, horizontal flipping. They used Adam optimizer and apply early stopping. Authors used Quadratic Weighted Kappa (QWK), Mean Absolute Error (MAE), F1 score and accuracy as performance metrics.

Results states that models using CDW-CE loss showed better performance. Also, visualization of Class Activation Maps results of the models traied with CDW-CE loss provide more accurate class discriminative regions which aligns with the medical experts. [1]

B. Diagnosis of ulcerative colitis from endoscopic images based on deep learning

This paper uses convolutional networks and recurrent neural network with an efficient attention mechanism to Ulcerative Collitis severity classification problem. Authors introduced a new neural network architecture called "UC-DenseNet" which uses "Efficient Attention Mechanism Network (EAM-Net)" along with "DenseNet201" and "Independently Recurrent Neural Networks (IndRNN)". DenseNet201 is used as the primary feature extractor. It is the convolutional backbone of the network.It comprises of five dense layers. Independently recurrent neural networks handles the temporal dependencies. By using the serialized data it can capture dynamic information for locating infected regions. This network also help model to avoid gradient explosion or vanishing. Efficient Channel Attention etwork (ECA-Net) uses 1D convolutions to assign importance weights to each feature channel which improves feature extraction. Convolutional Block Attention Module (CBAM) combines max pooling and global average pooling to construct spatial attention map. This spatial map multiplied with the original feature map. This way, it can locate spatial information. ECA-Net and CBAM constructs Efficient attention mechanism network (EAM-Net).

Extracted features from DenseNet201 are divided in two branches which are passed to IndRNN and EAM-Net. Attention maps from both modules are combined and used to highliht important features. For background feature retaining, global average pooling is used. The output is determined based on the pooled layers.

Authors used two datasets for evaluation of this task. The first dataset contains 9928 images obtained using the Olympus endoscopy system whereas the other dataset contains 4378 images using Fujinon endoscopy system.

Authors also used test time augmentation which are flipping and rotating the test data to improve robustness.

Authors used accuracy, precision, recall, F1 score and AUC as performance metrics. They have tested DenseNet201, Inception V3, ResNet152, VGG19, UC-DenseNet, DenseNet201+IndRnn+ECA-Net, DenseNet201+IndRnn+SE-Net, DenseNet201+IndRnn+CBAM. M-Net outperforms other attention mechanisms, such as SE-Net and CBAM. [2]

IV. Baseline Methods

A. Deep Learning Based Method

For the baseline model ResNet18 is selected as in [1].

a) ResNet18 architecture: ResNet18 is selected because it uses efficient residual blocks, it is computationally efficient. This architecture can effectively capture discriminative features in the dataset. Adam optimizer is used. Learning rate scheduler with scaling factor of 0.2 is used if validation set accuracy is not increased in last 10 epochs. Early stopping is applied also when training accuracy did not increase in the last 15 epochs. Models are trained at most for 100 epochs.

The model has benn trained with cross entropy loss function. Quadratic weight kappa (QWK), confusion matrix and accuracy is used as the performance metrics for this task.

b) Preprocessing Steps: Horizontal flipping and random rotation (-180-180) is applied to the input data. Also, input data are normalized using channel mean and standard deviation. By using horizontal flipping, new image variations are created which increases the training data diversity. Random rotation leads to better generalization and robustness. Channel normalization is applied to ensure that all input data have a consistent mean and standard deviation. By this way faster and stable convergence can be achieved during training.

c) Loss Function: Cross entropy loss function is used in this baseline model. Cross-entropy loss function quantifies the distance between the actual and the predicted classes. Its aim is to minimize the difference so that predicted probabilities are close to the true labels.

$$L = -\sum_{i=1}^{N} y_i \cdot \log(p_i) \tag{2}$$

This loss penalizes the incorrect predictions more which force the model to make accurate predictions.

d) Evaluation Metrics: Quadratic weight kappa and confusion matrix is used as evaluation metrics.

Quadratic weight kappa is designed for ordered classes and penalizes the predictions by calculating the distance from the actual class which makes it suitable for ordinal classification tasks. It penalizes wrong predictions of distant classes more than predictions of nearby classes. It is best suitable for imbalanced and ordial datasets.

Confusion matrix is a tabular summary which uses true positive, true negative, false positive and false negative. It allows for calculation of performance metrics such as recall, precision, F1 score which might give insight about the model performance.

e) Hyperparameter Tuning and Results: For obtaining the best performance of the model, hyperparameter tuning is applied to the model. Learning rate and weight decay is sweeped and logs are created using Weights&Biases (WANDB). Hyper-parameters for learning rate and weight decay are as follows: Learning rate 0.001, 0.0005, 0.0002 and Weight decay: 0, 0.001, 0.01. All combinations have been tried and the best performance is obtained using learning rate = 0.001 and weight decay = 0.001.

The results of the experiments can be found in Table 1. Confusion matrix can be found in Fig. 2 To observe the whole results which invludes all test scores, confusion matrices, validation accuracy, qwk score, loss and train accuracy, loss and qwk score please refer this wandb link.

TABLE I
Experiment Results for top 4 QWK scores.

	QWK	***Accuracy***
lr=0.001, wd=0.001	85.88%	78.29%
Lr=0.0005, wd=0.001	85.82%	78.41%
Lr=0.0002, wd=0.01	85.63%	77.22%
Lr=0.0005 wd=0.01	85.36%	77.11%

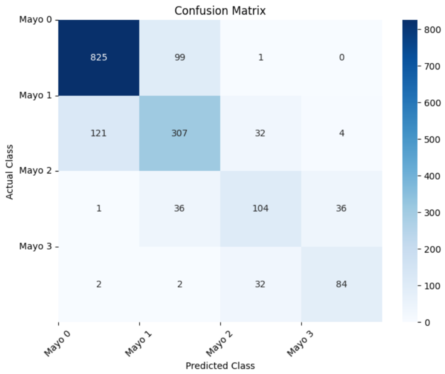

Fig. 2. Baseline Method Result

B. Naive Baseline

For naive baseline approach, stratified random baseline is used. This approach makes predictions based on the distribution of each class in the training data.Predictions are generated randomly however distributions are based on the original dataset distribution. This way, prediction pattern is more representative. This model aims to predict 'Mayo 0' for 54.14%, 'Mayo 1' for 27.7%, 'Mayo 2' for 11.12% and 'Mayo 3' for 7.67%. It is better model than pure random predictions and it is best suitable candidate for baseline method. Confusion matrix result for naive approach can be found in Fig. 3. QWK score of the baseline method is 0.017%.

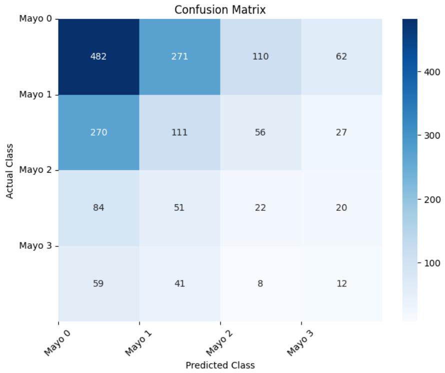

Fig. 3. Baseline Method Result

V. Transformer Model and Improvements

A. MaxVit

MaxVit is a hybrid architecture that combines the CNN's and transformers best benefits. It is vision transformer that can capture both local and global visual parts through multi-axial atteniton mechanism. Initial layers of MaxVit consist of CNN blocks which is responsible for effectively capturing the local features. To capture global dependencies in the images, transformer blocks are followed after the CNN blocks. The main advantage using MaxVit is the multi-axial attention mechanism which is a combination of local and axial attention. Axial attention is responsible for capturing long range dependencies along one dimension by operating along the height and width axes of the features maps. This operation is done independently which is computationally efficient. On the other hand, local attention captures the local dependencies within a small region to ensure details of the images are preserved. In this paper, pretrained version of MaxVit is used. ImageNet21k trained version is used, to ensure that model can generalize better and it will be more robust. Paper can be found [3].

B. Preprocessing Steps

Resizing, horizontal flipping and random rotation (-180-180) is applied to the input data. Also, input data are normalized using channel mean and standard devia- tion. By using horizontal flipping, new image variations are created which increases the training data diversity. Random rotation leads to better generalization and robustness. Channel normalization is applied to ensure that all input data have a consistent mean and standard deviation. By this way faster and stable convergence can be achieved during training. Also, MaxVit model accepts 224x224 input sized images, so resizing is applied.

C. Loss Function and Evaluation Metrics

Cross entropy loss function is used as main loss function and Quadratic Weighted Kappa (QWK) and confusion matrix is used as main evaluation metrics. Details of the loss function and evaluation metrics are explained in Baseline Methods section.

D. Hyperparameter Tuning and Results

For training the MaxVit model, batch size of 16 and AdamW optimizer is used. For hyperparameter tuning, learning rate values are changed. Tried hyperparameters are: 0.0001, 0.0002 and 0.0005. Models are trained for 50 epochs and if the validation accuracy is not increased for 10 epochs, early stopping is applied. The best performed model is when learning rate is 0.0001. Best qwk score is achieved as 88%. Confusion matrix can be found in Fig 4. Hyperparameter tuning results can be found in this wandb link.

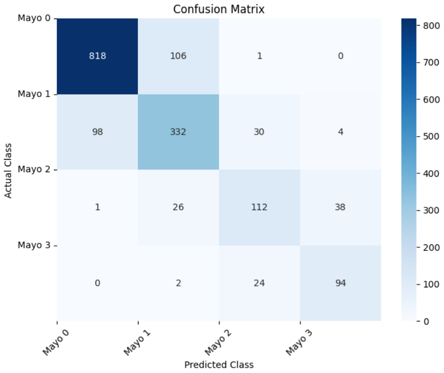

Fig. 4. Baseline Method Result

VI. Comparison and Discussion

MaxVit outperforms the baseline model performance as it can be seen in Table 2. The hybrid arcihtecture of MaxVit allows it to capture global and local variables. This makes feature extraction more efficient. The multi-axial attention mechanism makes a balance between computational efficiency and capturing lon range dependencies. This feature helps model to handle high resolution images and complex patterns without too much overhead. Since MaxVit is pretrained using ImageNet21k model can generalize better on new datasets.

ResNet18 on the other hand is a shallow traditional CNN model. It can capture local features more effectively but when it comes to long range dependencies, it performs worse.

TABLE II
Comaprison of Baseline Methods and MaxVit

	QWK	***Accuracy***
ResNet18	85.88%	78.29%
Naive Baseline	0.017%	-
MaxVit	88%	80.43%

VII. Conclusion

In this study on Mayo endoscopic classification, I compared the performance of two different neural network architectures: ResNet18 and MaxViT and a naïve baseline approach which use startified random sampling. The goal was to classify endoscopic images according to the Mayo scoring system to identitfy the severity of ulcerative colitis. ResNet18, is the baseline model. It achieved a Quadratic Weighted Kappa (QWK) score of 85% on the test set. This performance shows the effectiveness of CNNs in capturing local features and patterns within medical images. ResNet18's architecture, with its residual connections, facilitates training deeper networks and eliminates the vanishing gradient problem, making it a reliable choice for many image classification tasks. However, to increase the performance I applied MaxViT model pre-trained on the ImageNet21k dataset. MaxViT achieved QWK score of 88% on the test set, outperforming ResNet18. This performance increase is achieved because of MaxViT's hybrid architecture, which combines the CNNs and transformers. The initial convolutional layers in MaxViT effectively capture local features, while the transformer layers excel in modeling long-range dependencies and global context within the images. In summary, while ResNet18 demonstrated solid performance, the MaxViT model proved to be superior for Mayo endoscopic classification.

VIII. Future Work

This study shows that hybrid arcghitecture transformers and CNN's can be a good candidate for Mayo endoscopic classifiaction task. However, the ordinal structure of the dataset needs a loss function that takes into account this ordinality. Performance can further increase if the loss function is changed.

References

[1] G. Polat, I. Ergenc, H. T. Kani, Y. O. Alahdab, O. Atug, and A. Temizel, ”Class Distance Weighted Cross-Entropy Loss for Ulcerative Colitis Severity Estimation,” 2202.05167v2. Available: arXiv:2202.05167v2.

[2] X. Luo, J. Zhang, Z. Li, and R. Yang, ”Diagnosis of Ulcerative Colitis from Endoscopic Images based on Deep Learning,” Biomedical Signal Processing and Control, vol. 73, 2022, doi: 10.1016/j.bspc.2021.103443.

[3] M. T. Hassani, A. Walton, J. Navon, A. N. Ford, T. Aila, and A. Gholami, ”MaxViT: Multi-Axis Vision Transformer,” arXiv preprint arXiv:2204.01697, 2022.