Mixture-of-Experts, Multi-Task Learning for Eye Disease Diagnosis

Shishir Roy

July 2025

1 Introduction

In this assignment, we implemented a Mixture-of-Experts (MoE), Multi-Task Learning (MLT) Model that jointly performs disease grading (classification tasks) and lesion segmentation (segmentation tasks) from retinal images. We also trained independent Single-Task Models to serve as baselines for comparison. The code and results for this assignment can be found at
github.com/yoR-rihsihS/Multi_Task_Learning_Example

2 Dataset

We used the Indian Diabetic Retinopathy Image Dataset (IDRiD) [5], which provides the following:

- retinal fundus images and ground truth disease grading for Retinopathy (5 classes) and Risk of Macular Edema (3 classes).
- retinal fundus images and annotated binary masks (microaneurysms, haemorrhages, hard exudates, soft exudates, optic disc).
- We did not use any external data for training.

Comments:

- The dataset for disease grading and lesion segmentation are separate, i.e. for a given image we have either classification ground truths or segmentation ground truths.
- Furthermore, some images in segmentation dataset does not have all lesion masks ground truth. In such cases, I assumed absence of ground truth mask implies all the pixels belong to the negative class.
- We have referred Diabetic Retinopath Grade as ”rg” and Diabetic Macular Edema Risk as ”mer”.
- Due to the small size of the IDRiD (only 400 images for classification and 50 for segmentation), we did not use a separate validation set. We instead trained on the full training set and evaluated on the test set. While this maximizes data utilization for training, it may slightly overestimate generalization performance.

3 Preprocessing

- The images were resized to $268 \times 178, 1424 \times 2144, 712 \times 1072$ for single-task classification, single-task segmentation and multi-task classification-segmentation model respectively.
- The images were normalized using mean and standard deviation of ImageNet as we are using pretrained ResNet [3] as backbone for all our models.
- Random horizontal flip, random rotation, random rescale, random translate, random contrast adjustment, random blur were applied to enhance model robustness and mitigate overfitting.

4 Model Architecture

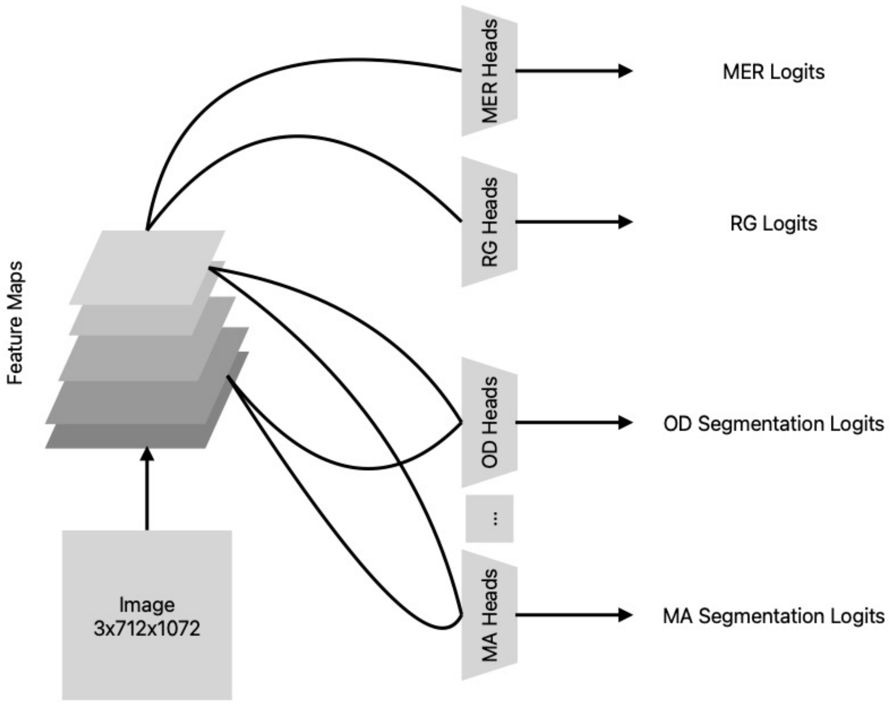
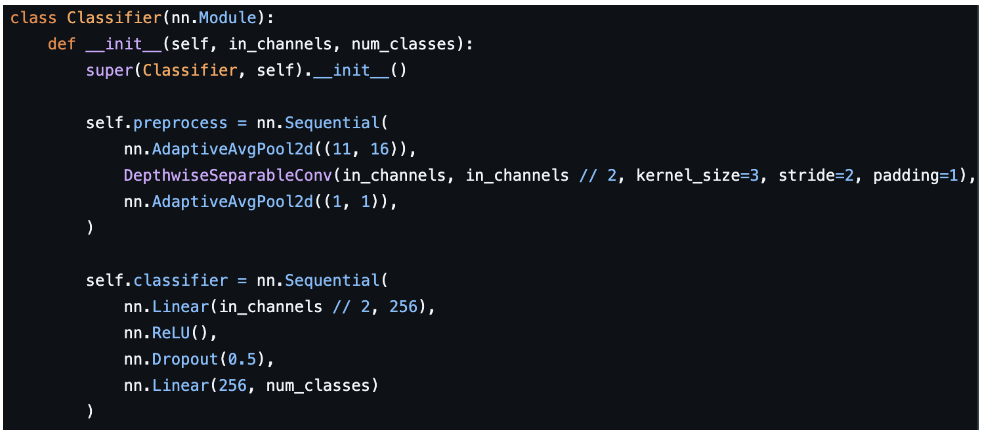


Figure 1: Architecture of our Mixture-of-Experts, Multi-Task Learning Model. Here, the final feature map of ResNet50 has output stride of 16.



```
class Classifier(nn.Module):
    def __init__(self, in_channels, num_classes):
        super(Classifier, self).__init__()

        self.preprocess = nn.Sequential(
            nn.AdaptiveAvgPool2d((11, 16)),
            DepthwiseSeparableConv(in_channels, in_channels // 2, kernel_size=3, stride=2, padding=1),
            nn.AdaptiveAvgPool2d((1, 1)),
        )

        self.classifier = nn.Sequential(
            nn.Linear(in_channels // 2, 256),
            nn.ReLU(),
            nn.Dropout(0.5),
            nn.Linear(256, num_classes)
        )
```

Figure 2: The spatial dimensions of final feature map is large, thus we first reduce the spatial dimension using AdaptiveAvgPool2d. Followed by a Depthwise Separable Convolution to learn features appropriate for classification and to reduce number of channels.

Segmentation Heads (ASPP Block and Decoder Block) are same as that in DeeplabV3Plus [2], the only difference is that any Convolution Layer with kernel size > 1 is swapped with Depthwise Separable Convolution. The use of Depthwise Separable Convolution [1] is made to control the number of parameters and Multiplication and Addition Operations, as we are employing Mixture-of-Expert paradigm.
The model in the beginning uses all the experts, so that they all can learn. After the model is trained for 50% of the epochs, we change this such that we select the most confident experts such that their combined confidence score is at least p [4]. This scheme is also used during inference time, which could make the model efficient.

5 Training and Evaluation Strategy

- We used a combination of loss functions to achieve robust training, cross entropy loss and focal loss were used for the classification task, while focal loss and dice loss were used for segmentation to better handle class imbalance.

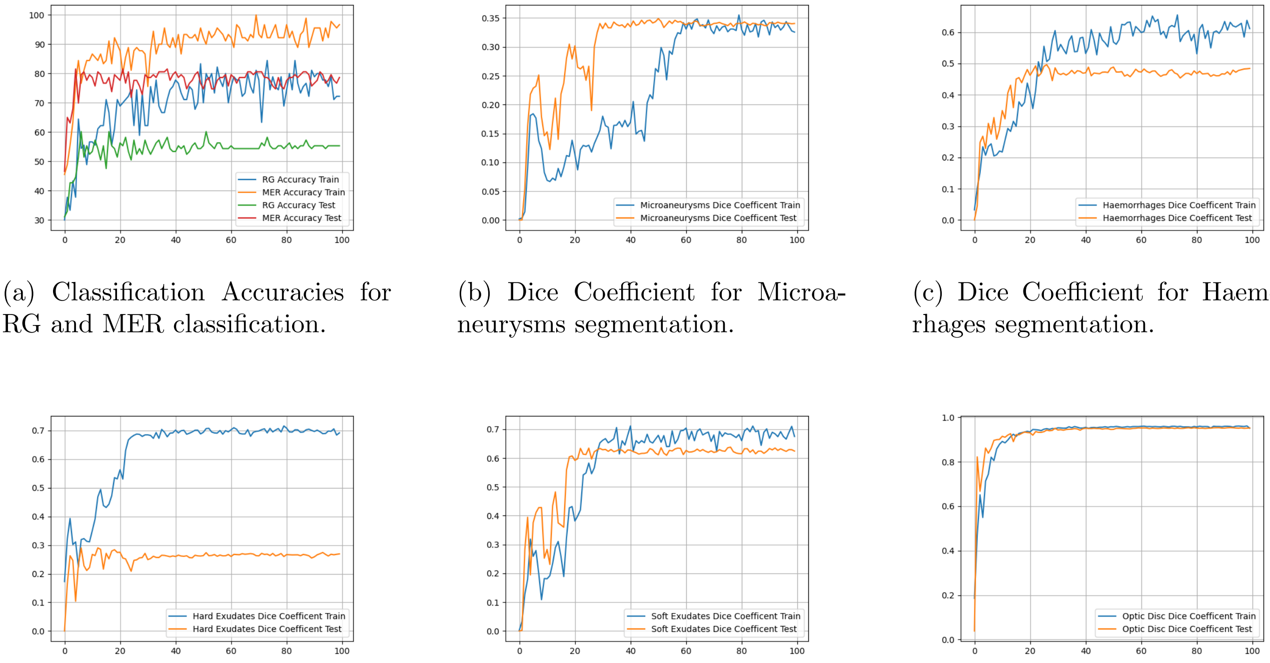
- The models were optimized using the AdamW optimizer, with an initial learning rate of 0.00003 for backbone parameters and 0.0003 for the remaining parameters. We also applied a weight decay of 0.0001 to mitigate overfitting and improve generalization.

- For evaluation, classification accuracy was computed as the percentage of correctly predicted samples. In the segmentation task, we computed IoU and dice coefficient, explicitly excluding background pixels to focus on the foreground regions of interest.

- However, while computing the segmentation loss during training, we included background pixels to ensure the model learns to separate pixels of interest from the background.

- We also optimize over some auxiliary losses for gating mechanism; namely l1 loss to encourage sparsity among the experts, (negative) entropy to discourage model from selecting same expert all the time.

6 Performance Metrics and Analysis



(a) Classification Accuracies for RG and MER classification.

(b) Dice Coefficient for Microaneurysms segmentation.

(c) Dice Coefficient for Haemorrhages segmentation.

(d) Dice Coefficient for Hard Exudates segmentation.

(e) Dice Coefficient for Soft Exudates segmentation.

(f) Dice Coefficient for Optic Disc segmentation.

Figure 3: Various performance metrics of Multi-Task Model during the course of training. To view such plots of other metrics and models visit the GitHub Repository of this project.

Model	Classification Accuracy		Segmentation IoU/Dice Score				
	RG	MER	MA	HE	EX	SE	OD
ResNet-50	54.3	81.5	-	-	-	-	-
DeepLabV3+	-	-	0.25/0.41	0.31/0.47	0.16/0.27	0.44/0.61	0.91/0.95
Multi-Task	55.3	78.6	0.20/0.34	0.31/0.48	0.15/0.26	0.45/0.62	0.90/0.95

Table 1: Performance Comparison of ResNet-50, DeepLabV3+ (ResNet-50 backbone) and Multi-Task Model (ResNet-50 backbone) with 3 experts per task.

7 Visualizations of Results

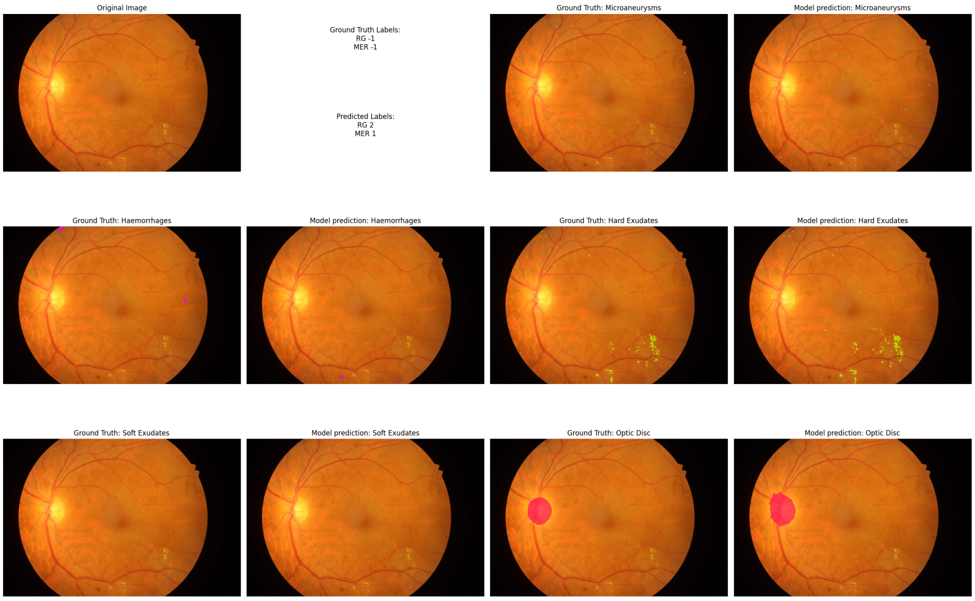

Original Image
Ground Truth Labels:
RG -1
MER -1
Predicted Labels:
RG 2
MER 1
Ground Truth: Microaneurysms
Model prediction: Microaneurysms
Ground Truth: Haemorrhages
Model prediction: Haemorrhages
Ground Truth: Hard Exudates
Model prediction: Hard Exudates
Ground Truth: Soft Exudates
Model prediction: Soft Exudates
Ground Truth: Optic Disc
Model prediction: Optic Disc

Figure 4: Model predictions visualized against ground truths. For this particular example classification ground truths are not available (-1).

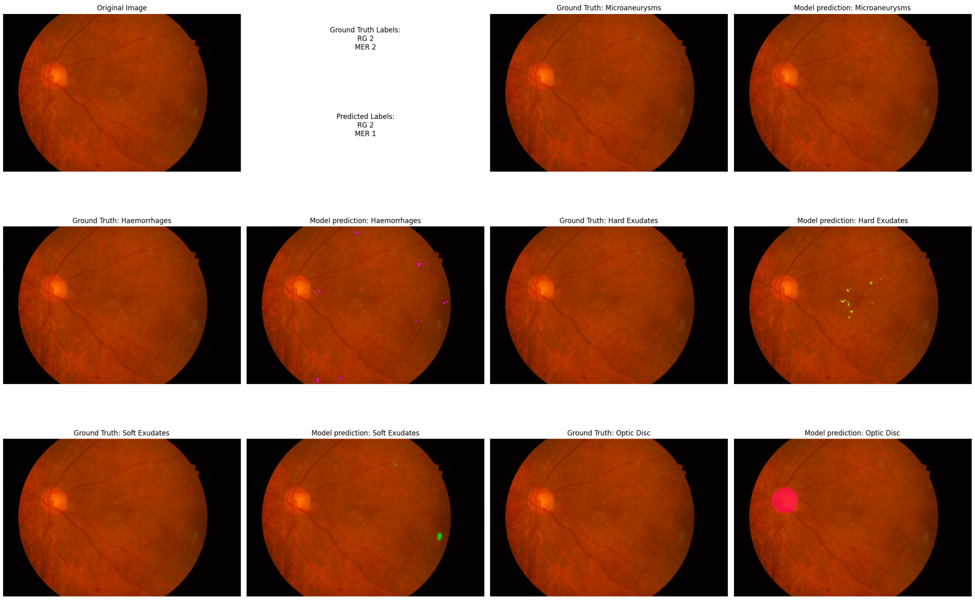

Original Image
Ground Truth Labels:
RG 2
MER 2
Predicted Labels:
RG 2
MER 1
Ground Truth: Microaneurysms
Model prediction: Microaneurysms
Ground Truth: Haemorrhages
Model prediction: Haemorrhages
Ground Truth: Hard Exudates
Model prediction: Hard Exudates
Ground Truth: Soft Exudates
Model prediction: Soft Exudates
Ground Truth: Optic Disc
Model prediction: Optic Disc

Figure 5: Model predictions visualized against ground truths. For this particular example segmentation ground truths are not available, thus original image is shown without any overlaid mask.

8 Conclusion

This assignment demonstrates the feasibility of a Mixture-of-Experts, Multi-Task Model for classification and segmentation. The model supports dynamic routing among experts based on the input, reducing resource requirements, and provides comprehensive outputs from a single forward pass.

References

[1] François Chollet. *Xception: Deep Learning with Depthwise Separable Convolutions.* IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[2] Liang-Chieh Chen and Yukun Zhu and George Papandreou and Florian Schroff and Hartwig Adam. *Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation.* European Conference on Computer Vision (ECCV), 2018.

[3] Kaiming He and Xiangyu Zhang and Shaoqing Ren and Jian Sun. *Deep Residual Learning for Image Recognition.* IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[4] Quzhe Huang and Zhenwei An and Nan Zhuang and Mingxu Tao and Chen Zhang and Yang Jin and Kun Xu and Kun Xu and Liwei Chen and Songfang Huang and Yansong Feng. *Harder Tasks Need More Experts: Dynamic Routing in MoE Models.* Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2024.

[5] Prasanna Porwal, Samiksha Pachade, Ravi Kamble, Manesh Kokare, Girish Deshmukh, Vivek Sahasrabuddhe, Fabrice Meriaudeau. *Indian Diabetic Retinopathy Image Dataset.* IEEE Dataport, 2018.