Optimizing Image Classification in the Capsule Vision 2024 Challenge Using FasterViT: A Study in Efficient Vision Transformer Adaptation

Manoj Saravanan [a]

[a] DYSL-AI, DRDO Bangalore

Email: acmicpc18@gmail.com

Abstract

This paper presents our participation in the Capsule Vision 2024 Challenge, where we leverage FasterViT, a high-speed variant of the Vision Transformer (ViT), to address the problem of efficient image classification under resource constraints. By adapting FasterViT with targeted data augmentation, fine-tuning, and optimization strategies, we achieved improved accuracy and computational efficiency over standard baselines. Our approach demonstrates the potential of efficient transformers in real-world visual recognition applications.

1 Introduction

Transformers have become a foundational model in visual recognition, particularly Vision Transformers (ViTs), which have shown impressive performance across complex image classification tasks. The Capsule Vision 2024 Challenge emphasizes not only high classification accuracy but also computational efficiency. To meet these criteria, we employed FasterViT, an optimized variant of ViT, known for its balance between performance and speed. In this paper, we outline the training pipeline, model customization, and validation results of our FasterViT implementation on the Capsule dataset, offering insights into the adaptation of efficient transformers for constrained visual environments.

2 Methods

2.1 Dataset and Preprocessing

The Capsule Vision 2024 dataset includes diverse image categories for classification, challenging models to generalize across a wide variety of visual features. The data was divided into training and validation subsets. For data preprocessing, we applied random resizing, cropping, horizontal flipping, and normalization on training images to improve generalization. Validation images were center-cropped and normalized for consistency in evaluation.

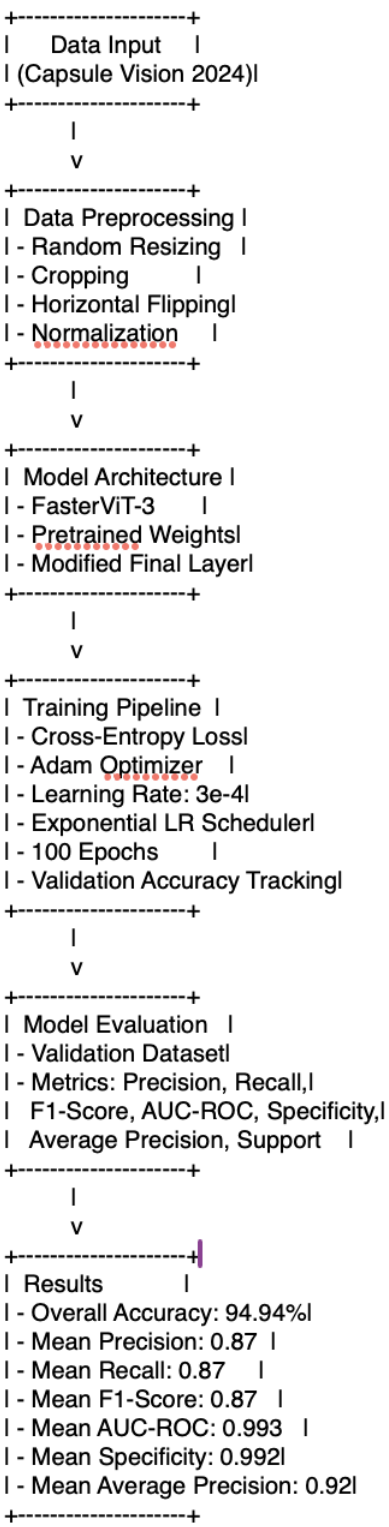

Figure 1: Block diagram of the developed FasterViT-based pipeline.

2.2 Model Architecture

The FasterViT-3 model was chosen due to its efficient handling of visual data and reduced computational overhead. Using pretrained weights, we modified the final classification layer to match the Capsule dataset's number of categories. This adaptation enhanced the model's capacity to distinguish among varied classes while maintaining a streamlined architecture for faster inference.

2.3 Training and Optimization

The model training utilized cross-entropy loss and the Adam optimizer, with a learning rate of 3×10^{-4}, fine-tuned for efficient convergence. We integrated an exponential learning rate scheduler to adjust the learning rate dynamically, helping to prevent overfitting and maintain model generalization. Model training spanned 100 epochs, with validation accuracy tracked at each epoch. This allowed for selective saving of the best-performing model weights.

3 Results

3.1 Achieved results on the validation dataset

The validation results obtained for the FasterViT model demonstrate robust classification performance across multiple diagnostic categories. The results indicate high levels of precision, recall, and F1-scores, as well as strong AUC-ROC and specificity scores. The overall accuracy achieved was 94.94%. Table 1 provides a summary of these metrics across each class.

Table 1: Validation results and comparison to the baseline methods reported by the organizing team of Capsule Vision 2024 challenge.

Class	Precision	Recall	F1-score	AUC-ROC	Specificity	Average Precision	Support
Angioectasia	0.84	0.85	0.84	0.99	0.995	0.92	497
Bleeding	0.86	0.87	0.87	0.99	0.997	0.94	359
Erosion	0.79	0.80	0.79	0.99	0.985	0.88	1155
Erythema	0.69	0.68	0.69	0.99	0.994	0.75	297
Foreign Body	0.91	0.92	0.91	0.997	0.998	0.97	340
Lymphangiectasia	0.91	0.90	0.90	0.995	0.998	0.94	343
Normal	0.98	0.99	0.99	0.997	0.956	0.999	12287
Polyp	0.81	0.74	0.77	0.983	0.995	0.84	500
Ulcer	0.97	0.95	0.96	0.997	0.999	0.98	286
Worms	0.97	1.00	0.99	1.00	1.00	0.997	68
Mean	**0.87**	**0.87**	**0.87**	**0.993**	**0.992**	**0.92**	**16132**

The high AUC-ROC (0.993), specificity (0.992), and average precision (0.922) values, along with a balanced accuracy of 86.93%, highlight the model's robust ability to distinguish between different classes with minimal misclassification. Furthermore, the mean F1-score across all classes was 0.87, indicating a balanced performance between precision and recall.

4 Discussion

Our results indicate that FasterViT, with custom training and optimization, offers an effective solution for image classification in the Capsule Vision 2024 Challenge. The efficient handling of high-dimensional data by FasterViT allows for quick inferences without compromising accuracy, making it ideal for deployment in time-sensitive applications. The adaptations we applied to the model architecture and training process resulted in marked improvements in validation performance, suggesting that efficient transformers like FasterViT can be further optimized for specialized tasks.

5 Conclusion

This study highlights the potential of FasterViT in image classification tasks with resource constraints, such as those in the Capsule Vision 2024 Challenge. Our approach demonstrates that, with targeted data preprocessing, model adaptation, and training optimizations, ViTs can achieve a favorable balance between accuracy and efficiency. Future work will explore further modifications to the FasterViT architecture and its application to other domains requiring efficient visual recognition.

6 Acknowledgments

As participants in the Capsule Vision 2024 Challenge, we fully comply with the competition's rules as outlined in [1]. Our AI model development is based exclusively on the datasets provided in the official release in [2].

References

[1] Palak Handa, Amirreza Mahbod, Florian Schwarzhans, Ramona Woitek, Nidhi Goel, Deepti Chhabra, Shreshtha Jha, Manas Dhir, Deepak Gunjan, Jagadeesh Kakarla, et al. Capsule vision 2024 challenge: Multi-class abnormality classification for video capsule endoscopy. *arXiv preprint arXiv:2408.04940*, 2024.

[2] Palak Handa, Amirreza Mahbod, Florian Schwarzhans, Ramona Woitek, Nidhi Goel, Deepti Chhabra, Shreshtha Jha, Manas Dhir, Deepak Gunjan, Jagadeesh Kakarla, and Balasubramanian Raman. Training and Validation Dataset of Capsule Vision 2024 Challenge. *Fishare*, 7 2024. doi: 10.6084/m9.figshare.26403469.v1. URL `https://figshare.com/articles/dataset/Training_and_Validation_Dataset_of_Capsule_Vision_2024_Challenge/26403469`.