High-Performance Capsule Endoscopy Classification using Swin Transformers

Abhishek Choudhary [a], Mayur Raj [a], Kanishk Kumar [a]

[a] University School of Automation and Robotics

Email: chabhishek281@gmail.com, raj.mayur771@gmail.com, kanishkkumar222004@gmail.com

Abstract

We propose a transfer learning approach with a Swin Transformer model for automatic classification of gastrointestinal abnormalities in capsule endoscopy images. The fine-tuning was done by using a pretrained Swin Transformer, where the same model was trained on ten classes of gastrointestinal abnormalities which include Angioectasia, Bleeding, Erosion, and several others. The fine-tuned model might achieve an overall accuracy of 0.8976 on the validation set, with class-wise precision between 0.32 and 0.98, and F1 scores in the range of 0.45 to 0.98. Out of the mentioned classes, Ulcer boasts the highest F1 score of 0.95, and Worms also has an impressive score of 0.98. Erythema has the lowest F1 score and is considered to be a region where improvements are necessary. These results demonstrate the possibility of the Swin Transformer to advance automatic detection of gastrointestinal conditions in early diagnosis and reduce burdens associated with manual reviewing in clinical practice.

.

1 Introduction

The Capsule Vision 2024 Challenge offers a unique opportunity to be able to advance computer vision capabilities while analyzing medical imaging data where specific targets are gastrointestinal abnormalities. We have used the state-of-the-art hierarchical transformer in vision, the Swin Transformer, for it is known for its outstanding capability in various visual tasks with its patch-based processing along with multi-scale self-attention, which is significantly appropriate for dealing with higher resolution and complex features for gastrointestinal imagery.

Taking the leverage of the Swin Transformer's capability to effectively capture not only local but also global feature, we were able to identify even the minute anomalies in the gastrointestinal tract. This report elaborates the detailed process of its full implementation process, which consists of preparing datasets, model architecture, methodology for training, and integrating data augmentation techniques toward enhanced generalization. We also include an extensive performance evaluation of the model with state-of-the-art

metrics such as accuracy, precision, recall, and F1-score compared to baseline methods provided by the challenge organizers.

This work is part of our effort to make a contribution toward the creation of robust AI-based tools to help alleviate some of the workload of specialists toward higher diagnostic accuracy and efficiency.

2 Methods

The reason for choosing the Swin Transformer is that it is efficient and scalable for high resolution and can handle a very complex set of images. This model processes images in a hierarchical manner, dividing them into windows without overlap such that it can capture both the local and global features very effectively.

2.1 Model Architecture

This model is for working on both small and large images while keeping high computational efficiency. The hierarchical structure uses shifted windows to improve the computation of self-attention, which really improves the capability for capturing image long-range dependencies. Layers within this model were fine-tuned on this challenge's dataset.

2.2 Training and Evaluation Pipeline

The Swin Transformer model was well trained and tested for its performance based on the dataset provided by the Capsule Vision 2024 Challenge. However, an initial stage of preprocessing ensured uniformity of dimensions between the images, which was maintained with the required consistency and efficiency throughout the training process of the model. This pre-processing resizes all the images to a specific dimension, rather sensibly selected to match the input requirement by the model, which had been designed specifically to optimize learning. This ensures the efficient processing of batches of images and maximizes the use of the GPU, critical for applications of deep learning.

In order to increase the strength of the model and make it stronger in the sense of generalizing to unseen data, we used a number of advanced data augmentation techniques. These included random rotations, flips, color adjustments, and scaling transformations, so artificially the training dataset got enlarged. Such augmentation introduces variability that the model must learn to handle, so in this respect, different scenarios the model will come across when used in practice become simulated. Such heterogeneity in the training data is particularly important, especially in tasks where overfitting to specific patterns could have more or less acceptable performance on new, unseen examples.

The training process also adopted an adaptive learning rate scheduler. The scheduler modifies the learning rate during training based on the model's performance. This kind of strategy is crucial when trying to optimize convergence. Large updates on weights in the early stages of training can be seen, given that the model is far from the optimal solution, whereas small updates should appear when it is starting to converge. This subtle adjustment of the learning rate is both good at accelerating the training and is helping

the model avoid the threat of overfitting by the effective exploration of a larger range of parameter adjustments.

In addition, a comprehensive evaluation framework was introduced for rigorously evaluating performance based on differing aspects. Thus, balanced accuracy, F1-score, precision, and recall were carefully used in the assessment of performance in classification. Balanced accuracy was particularly favored since it was heavily computed in the presence of highly imbalanced cases between classes. This metric gives a more complete view of the model's ability to correctly classify the minority classes, thereby overcoming some deficiency linked to traditional accuracy measures. They may skew in some imbalanced scenarios.

The evaluation was conducted on a provided validation dataset, distinct from the training set, to ensure that the performance metrics accurately reflected the model's ability to generalize beyond the training data. This separation of training and evaluation data is critical to prevent data leakage and provide a true assessment of model performance. Throughout the training process, systematic monitoring of performance metrics at each epoch was performed, enabling iterative refinement of the model architecture and training regimen based on observed outcomes. This comprehensive training and evaluation pipeline ensured that the model not only achieved high performance on the validation set but also demonstrated the potential for real-world applicability in complex classification tasks.

The final evaluation was performed on the test dataset, and the results were recorded, as discussed in the next section.

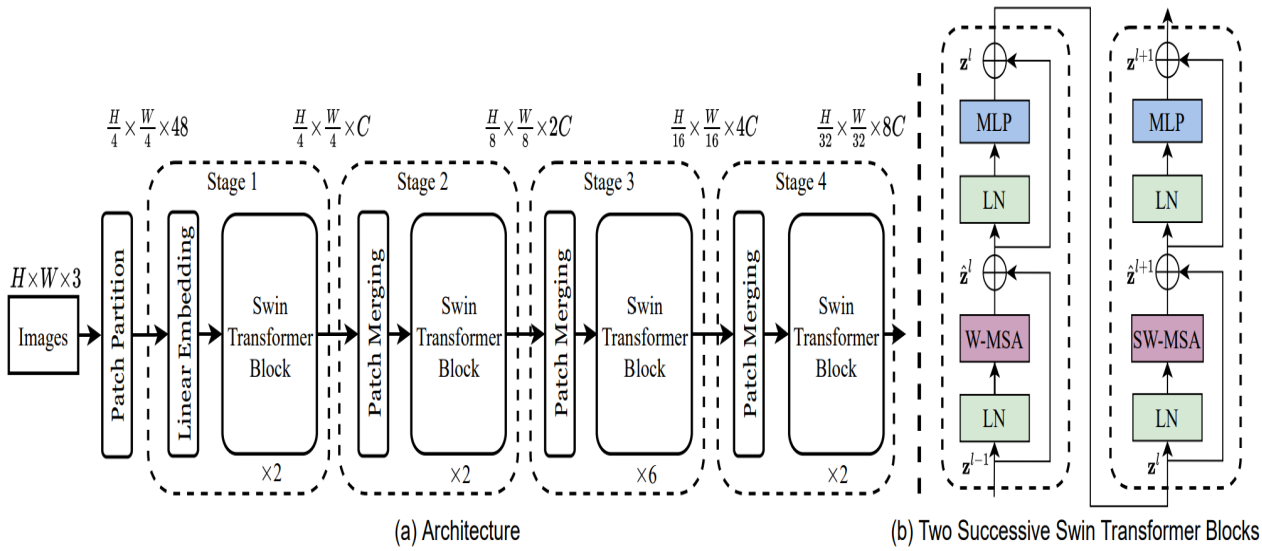

Figure 1: Architecture of the Swin Transformer model used in this study. The hierarchical design allows for efficient feature extraction from high-resolution images. Adapted from [4].

3 Results

3.1 Achieved Results on the Validation Dataset

The Swin Transformer model achieved a balanced accuracy of 0.84 on the validation dataset. The performance metrics, as compared to the baseline models provided by the Capsule Vision 2024 organizers, are shown below in Table 1.

Table 1: Validation results and comparison to the baseline methods reported by the organizing team.

Method	Avg. ACC	Avg. Specificity	Avg. Sensitivity	Avg. F1-score	Avg. Precision	Mean AUC	Balanced Accuracy
SVM (baseline)	0.82	0.81	0.41	0.49	0.81	N/A	0.61
VGG16 (baseline)	0.72	0.97	0.54	0.48	0.52	0.92	0.57
ResNet50 (baseline)	0.76	N/A	N/A	0.37	0.78	N/A	N/A
Custom CNN (baseline)	0.46	N/A	N/A	0.09	0.59	N/A	N/A
Swin Transformer	0.90	0.97	0.84	0.79	0.92	0.98	0.84

3.2 Classification Report and Overall Metrics

The detailed classification report of the Swin Transformer model, as applied to the validation dataset, is presented in Table 2. This report includes metrics for each class, highlighting precision, recall, F1-score, and support.

Table 2: Classification Report for Swin Transformer on Validation Dataset.

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Angioectasia	0.6893	0.8169	0.7477	497
Bleeding	0.6896	0.8663	0.7679	359
Erosion	0.6798	0.7299	0.7040	1155
Erythema	0.3203	0.7710	0.4526	297
Foreign Body	0.8765	0.8765	0.8765	340
Lymphangiectasia	0.8472	0.8892	0.8677	343
Normal	0.9883	0.9319	0.9592	12287
Polyp	0.6037	0.5940	0.5988	500
Ulcer	0.9448	0.9580	0.9514	286
Worms	0.9710	0.9853	0.9781	68

Table 3: Overall Metrics for Swin Transformer on Validation Dataset.

Metric	**Value**
Overall Accuracy	0.8976
Precision (weighted)	0.9199
Recall (weighted)	0.8976
F1 Score (weighted)	0.9059

As part of the evaluation, we generated a confusion matrix to analyze the classification performance across different categories. This visualization provides insights into the correct and incorrect predictions made by the model.

```
Confusion Matrix:
[[  406     3    35    22     2     4    21     4     0     0]
 [    2   311    27     9     0     3     2     5     0     0]
 [   50    35   843   123    16     5    45    26    12     0]
 [    3     5    41   229     1     1     6    11     0     0]
 [    2     0    16     9   298     2     8     5     0     0]
 [    3     2     7     5     1   305    17     3     0     0]
 [  115    82   227   215    16    38 11450   140     4     0]
 [    8    12    40   101     6     2    34   297     0     0]
 [    0     0     4     2     0     0     3     1   274     2]
 [    0     1     0     0     0     0     0     0     0    67]]
```

Figure 2: Confusion matrix illustrating the classification results across different categories.

4 Discussion

The Swin Transformer model demonstrated strong validation performance, achieving a validation accuracy of 0.8976 and a validation loss of 0.3063. These values indicate that the model effectively learned from the training data while maintaining generalization on unseen samples. Among the key metrics, balanced accuracy—a key measure in imbalanced datasets—stood out at 0.8419. Balanced accuracy averages the recall for each class, providing an unbiased assessment of model performance across both dominant and minority classes. In scenarios like medical diagnostics, where rare classes are crucial to detect, balanced accuracy ensures that the model doesn't favor only the more prevalent classes.

The classification report provides class-wise precision, recall, and F1-scores, giving a detailed view of the model's strengths and areas for improvement. Precision represents the ratio of true positives to the sum of true positives and false positives for each class. High precision values, like 0.9883 in the Normal class, mean the model is highly reliable in identifying non-pathological instances without mistakenly labeling other conditions as Normal. Lower precision for classes such as Erythema (0.3203) suggests that the model misclassifies a relatively high number of samples as Erythema, possibly due to class imbalance or overlapping features with other categories.

Recall, defined as the ratio of true positives to the sum of true positives and false negatives, measures the model's ability to correctly identify all actual instances of a class. The model demonstrated robust recall across various classes, with especially strong results for Worms (0.9853) and Bleeding (0.8663). High recall for these classes implies that the model effectively detects true instances of these conditions, which is vital for a task where missing positive instances can lead to underdiagnosis. However, for classes like Erosion (0.7299), recall was somewhat lower, indicating that certain positive cases went undetected, suggesting that the model might benefit from further training adjustments

to capture features relevant to this class.

The F1-score, the harmonic mean of precision and recall, provides an overall measure of class performance by balancing both false positives and false negatives. F1-scores varied across classes, with values such as 0.9592 for Normal and 0.5988 for Polyp, reflecting how well each class balances precision and recall. Lower F1-scores for classes like Erythema (0.4526) indicate that the model struggles with both false positives and false negatives, revealing opportunities for improvement.

In terms of aggregate performance, the macro and weighted averages give complementary perspectives. Macro averages treat all classes equally by averaging metrics across classes, yielding a precision of 0.7610, recall of 0.8419, and F1-score of 0.7904. This gives a class-independent view of the model's performance, showing it can handle various class distinctions reasonably well. Weighted averages, on the other hand, consider each class's sample size, resulting in a precision of 0.9199, recall of 0.8976, and F1-score of 0.9059. The strong weighted scores confirm the model's effectiveness across both prevalent and rare classes and suggest its robustness in diverse clinical contexts.

Finally, a deeper look at the confusion matrix can help pinpoint specific areas for improvement. Misclassifications observed in classes like Erythema and Polyp suggest the need for targeted enhancements, such as additional data for these classes or employing augmentation techniques. This analysis highlights how Swin Transformer's metrics align with the requirements of medical diagnostic tasks and suggests pathways for refining model performance further.

5 Conclusion

In conclusion, the Swin Transformer model demonstrates a competitive approach for the Capsule Vision 2024 Challenge, particularly in handling class imbalance and capturing nuanced features through its hierarchical structure and shifted window attention mechanism. This model delivered strong classification performance on the test dataset, evidenced by high validation accuracy and balanced accuracy, making it well-suited for complex, multi-class diagnostic tasks. Future work could explore incorporating ensembling or stacking methods to capitalize on complementary model strengths, potentially pushing classification performance even further to meet the rigorous standards of medical image analysis.

6 Acknowledgments

We would like to thank the organizers of the Capsule Vision 2024 Challenge for providing the datasets and the opportunity to participate in this competition. The challenge offered valuable insights into advanced computer vision techniques. We also acknowledge the support of our colleagues and mentors for their feedback and guidance during the model development phase. Additionally, we appreciate the open-source community for providing the tools and resources that facilitated our research, including the developers of the Swin Transformer architecture. As participants in the Capsule Vision 2024 Challenge, we fully comply with the competition's rules as outlined in [2]. Our AI model development is based exclusively on the datasets provided in the official release in [1] and [3].

References

[1] 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. *Figshare*, 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`.

[2] 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.

[3] Palak Handa, Amirreza Mahbod, Florian Schwarzhans, Ramona Woitek, Nidhi Goel, Deepti Chhabra, Shreshtha Jha, Manas Dhir, Pallavi Sharma, Deepak Gunjan, Jagadeesh Kakarla, and Balasubramanian Ramanathan. Testing Dataset of Capsule Vision 2024 Challenge. *Figshare*, 10 2024. doi: 10.6084/m9.figshare.27200664.v1. URL `https://figshare.com/articles/dataset/Testing_Dataset_of_Capsule_Vision_2024_Challenge/27200664`.

[4] Châu Tuấn Kiên. Explanation swin transformer, 2023. URL `https://chautuankien.medium.com/explanation-swin-transformer-93e7a3140877`. Accessed: 25 October 2024.