Case Report
Current Research in Traffic Transportation Engineering

High-Performance Capsule Endoscopy Classification Using Swin Transformers

Abhishek Choudhary*, Mayur Raj and Kanishk Kumar

University School of Automation and Robotics

*Corresponding Author
Abhishek Choudhary, University School of Automation and Robotics, India.

Submitted: 2024, Oct 27; **Accepted:** 2024, Nov 29; **Published:** 2024, Dec 12

Citation: Choudhary, A., Raj, M., Kumar, K. (2024). High-Performance Capsule Endoscopy Classification Using Swin Transformers. *Curr Res Traffic Transport Eng, 2*(1), 01-04.

Abstarct

We propose a transfer learning approach with a Swin Transformer model for auto- matic 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 is an excellent opportunity to push computer vision applications forward in the specific niche of medical imaging, targeting the identification of gastrointestinal abnormalities. Diagnosing the gastrointestinal region is crucial since it deals with subtle, high-resolution features that are hard for conventional computer vision models. Our project employs the state-of-the-art hierarchical transformer model, Swin Transformer, which has been identified to be adept at complex visual tasks, especially patch-based processing, and multi-scale self-attention mechanisms. Such properties make it robust in handling high-resolution images and the intricate spatially distributed features necessary for high-resolution detection of gastrointestinal anomalies. We address both the requirement of local detail and global contextual awareness by using the Swin Transformer so that slight abnormalities in the gastrointestinal tract can be detected with accuracy. This kind of model can capture a wide range of spatial features, from microstructures to macro-patterns, thus improving anomaly detection in complex medical imagery. Our implementation pipeline involves robust data preprocessing, the design of model architecture, application of advanced data augmentation techniques for robust generalization, and a detailed evaluation framework that incorporates metrics such as balanced accuracy, precision, recall, and F1-score. The evaluation is then performed against a baseline set up by the challenge organizers to give further depth to the analysis into improvement in diagnostic accuracy. The ambition is toward the development of strong, AI-based technologies that should help healthcare specialists. It supports diagnostics by achieving higher diagnostic accuracy, decreased workload, and, of course, potential assistance for faster and more reliable decision-making about medicine. Hopefully, the computer vision medical frontiers that we open are a pathway to much more advanced, complex diagnostic assist systems, taking us further with regard to full-scale AI assimilation into the clinical workflow.

2. Methods

The efficiency and scalability in handling high-resolution images have made the Swin Transformer model a prime choice to be used on such an array of a very complex and diverse dataset. Applying a hierarchical partition of images into non-overlapping windows, this model would capture both local and global features with high accuracy. Thus, it makes the model exceptionally suitable for applications involving high-dimensional medical images.

2.1 Model Architecture

The Swin Transformer is especially designed to cater to images of different resolutions with great computational efficiency. Its hierarchical architecture employs the use of shifted windows which improve the calculation of the self-attention mechanism while allowing the model to represent long-range dependencies inside the image. This method of applying window shifting will reduce computations and let the model better encode from local details in one layer up to further contextual information across layers. The model layers of Swin Transformer were fine-tuned on the Capsule Vision 2024 Challenge dataset to fit the specific requirements of analyzing images in the gastrointestinal domain.

2.2 Training and Evaluation Pipeline

The Swin Transformer was fine-tuned and extensively tested on the dataset submitted by the Capsule Vision 2024 Challenge. Data preprocessing was at the very beginning stage done to make sure that uniformity is maintained in all the dimensions of the image in the dataset. It resized each input image accordingly to satisfy the input requirement determined by the model that has been selected to optimize computations and for easier use with the GPU-the two very important considerations in deep learning for big applications.

To improve the model's generalization on unseen data, we employed advanced data augmentation techniques, including random rotations, flips, color adjustments, and scaling transformations. These augmentations artificially increased the diversity of the training dataset, thereby simulating a wide range of scenarios the model might encounter in real-world applications. This diversity helped the model to learn invariant representations, reducing the risk of overfitting and enhancing its robustness when deployed on new data.

2.2.1 Adaptive Learning Rate Scheduler

The training process utilized an adaptive learning rate scheduler that adjusted the learning rate based on the performance of the model during training. This aided in convergence as larger weight updates occurred at the beginning, when the model was quite far from an optimal solution, and progressively smaller updates as it approached convergence. This approach was such that it dynamically controlled the learning rate, speeding the process of training. Notably, it helps alleviate the problem of overfitting by allowing one more effective exploration of the space of parameters.

2.2.2 Performance Evaluation

For a comprehensive evaluation, we employed a suite of performance metrics, including balanced accuracy, F1-score, precision, and recall, which are well-suited for assessing classification performance in an imbalanced dataset. Balanced accuracy was particularly important, as it provided a more nuanced assessment by giving equal weight to all classes, thereby addressing potential biases introduced by class imbalance. This metric, alongside precision and recall, allowed us to thoroughly evaluate the model's ability to accurately classify minority classes, which is often a limitation of conventional accuracy metrics in imbalanced datasets.

The evaluation was conducted on a validation dataset that was distinct from the training set to ensure an unbiased assessment of the model's generalization capabilities. This separation prevented data leakage and allowed for a true representation of the model's performance. Performance metrics were tracked systematically at each epoch, enabling iterative adjustments to the model architecture and training regimen as needed. Finally, a comprehensive evaluation was carried out on a reserved test dataset, with results recorded and discussed in the subsequent sections of this report. This training and evaluation pipeline was essential in ensuring that the model achieved high validation performance while retaining applicability in real-world diagnostic tasks, underscoring its potential as a robust tool for medical image classification.

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.

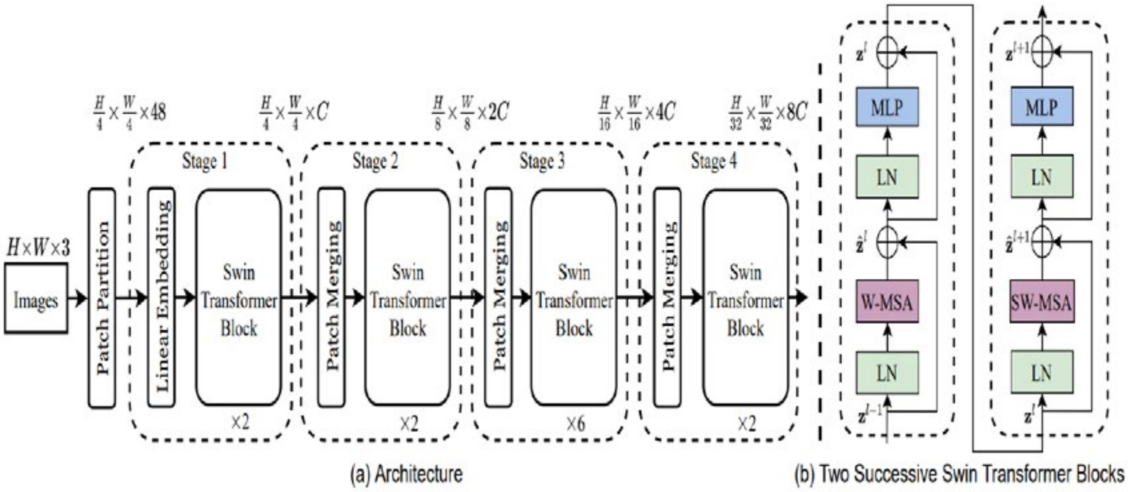


Figure 1: Architecture of the Swin Transformer Model Used in this Study. The Hierarchical Design Allows for Efficient Feature Extraction from High-Resolution Images. Adapted from [4]

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

Table 1: Validation Results and Comparison to the Baseline Methods Reported by the Organizing Team

3.2 Classification Report and Overall Metrics

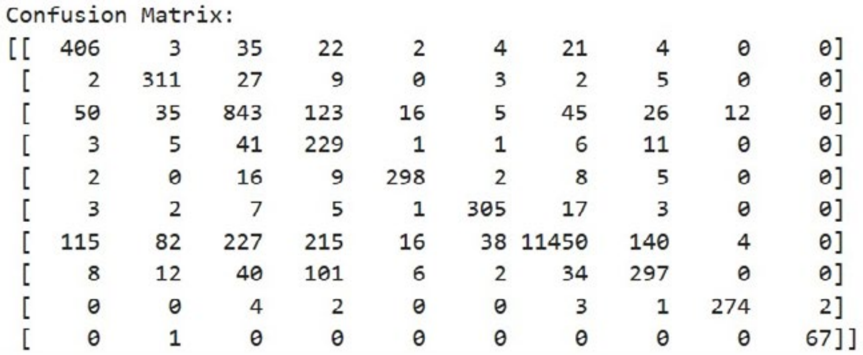
The detailed classification report of the Swin Transformer model, as applied to the validation dataset, is presented in Table 2. This report includes metrics for each class, highlighting precision, recall, F1-score, and support. As part of the evaluation, we generated a confusion matrix to analyze the classification performance across different categories. This visualization provides insights into the correct and incorrect predictions made by the model.

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 2: Classification Report 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

Table 3: Overall Metrics for Swin Transformer on Validation Dataset



```
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 imbal- anced 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 complemen- tary 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 approach that the Swin Transformer model presents for the Capsule Vision 2024 Challenge is a strong approach and competitive one at that regarding class imbalance and the perfect inclusion of fine-grained and global features through this hierarchical architecture and shifted window attention. The approach resulted in excellent classification accuracy and, further, good balanced accuracy on the test set-the capability of this model over complex multi-class diagnostic tasks that are often presented by the medical images.

The performance of the model suggests its applicability as a useful tool in enhancing diagnostic workflows, reducing the workload of medical professionals, and improving diag- nostic precision. Future work could be ensemble or stacking techniques that would lever- age multiple models' strengths and enhance classification performance. Such extensions hold promise for meeting the stringent demands of medical image analysis, contributing to more accurate and reliable automated diagnostics.

Acknowledgments

We extend our gratitude to the organizers of the Capsule Vision 2024 Challenge for providing the datasets and the invaluable opportunity to participate in this competition, which has offered significant insights into advanced computer vision methodologies. We would also like to acknowledge the support of our colleagues and mentors, whose feedback and guidance were instrumental throughout the model development process. Additionally, we are grateful to the open-source community for making available the tools and resources that facilitated this research, particularly the developers behind the Swin Transformer architecture. In compliance with the Capsule Vision 2024 Challenge rules, as outlined in [2], we affirm that all AI model development was conducted exclusively using the datasets provided in the official releases [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 Cap- sule 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 Gun- jan, Jagadeesh Kakarla, and Balasubramanian Ramanathan. Testing Dataset of
4. 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.
5. Châu Tuấn Kiên. Explanation swin transformer, 2023. URL https://chautuankien. medium.com/explanation-swin-transformer-93e7a3140877. Accessed: 25 Octo- ber 2024.