Multi-Class Abnormality Classification Task in Video Capsule Endoscopy

Dev Rishi Verma[1], Vibhor Saxena[1], Dhruv Sharma[1], Arpan Gupta[1,2]

[1] Institue of Engineering and Technology, JK Lakshmipat University, Jaipur, 302026 Rajasthan, India

[2]Email: arpan.gupta@jklu.edu.in

Abstract

In this work we addressed the challenge of multi-class anomaly classification in Video Capsule Endoscopy (VCE)[1] with a variety of deep learning models, ranging from custom CNNs to advanced transformer architectures. The purpose is to correctly classify diverse gastrointestinal disorders, which is critical for increasing diagnostic efficiency in clinical settings. We started with a proprietary CNN and improved performance with ResNet[7] for better feature extraction, followed by Vision Transformer (ViT)[2] to capture global dependencies. Multiscale Vision Transformer (MViT)[6] improved hierarchical feature extraction, while Dual Attention Vision Transformer (DaViT)[4] delivered cutting-edge results by combining spatial and channel attention methods. This methodology enabled us to improve model accuracy across a wide range of criteria, greatly surpassing older methods.

1 Introduction

The entire gastrointestinal (GI) tract, especially the small intestine, which is difficult to reach with traditional endoscopy, can be examined with Video Capsule Endoscopy (VCE)[1], a non-invasive diagnostic technique. However, manual analysis is time-consuming and prone to errors due to the enormous volume of video frames produced by every procedure. Deep learning-based automated systems have enormous potential to improve the precision and effectiveness of identifying gastrointestinal anomalies such as ulcers, polyps, and bleeding. Recent developments in transformer architectures have demonstrated incredible promise in this area, especially with models such as the Vision Transformer (ViT)[2], which use attention mechanisms and hierarchical processing of visual information to dramatically improve performance in image recognition tasks. [2].

In order to solve this issue, the Capsule Vision 2024 Challenge promotes the creation of sophisticated AI models that can classify abnormalities across multiple classes. In order to improve the classification performance, we investigate a variety of deep learning architectures in this report, starting with a custom CNN and working our way up to more complex models like the Vision Transformer (ViT) and its variations. By combining convolutional techniques with attention mechanisms, our method takes advantage of the spatial-temporal complexity of VCE data. This improves the model's capacity to identify both local and global dependencies within the video frames. Furthermore, prior research has shown that deep learning methods can successfully detect anomalies in VCE, which offers a strong basis for our DaViT model's development.

2 Methods

This section offers a detailed analysis of the architecture of the Dual Attention Vision Transformer (DaViT)[4], our best model for video capsule endoscopy (VCE) abnormality classification. Please visit our GitHub repository to get a detailed description of the several deep learning architectures we assessed, including our proprietary CNN and Vision Transformer versions.[8]

Channel Group Self-Attention

By using unique 'channel tokens' that consider each colour layer as a whole, this technique generates tokens based on image colour channels, aiding the model in comprehending the image as a whole. The outcome of this procedure is displayed in the final output representation Z as follows:

$$Z = \sum_{c=1}^{C} \alpha_c \cdot f_c(X) \tag{1}$$

where:

- Z is the output representation.
- α_c is the attention weight of channel c.
- $f_c(X)$ represents the feature extraction function for the channel.

Spatial Window Multihead Self-Attention

This technique divides the spatial dimension into local windows and uses a multihead approach to handle numerous spatial tokens. The following is a representation of the attention computation:

$$A_s = \text{softmax}\left(\frac{Q_s K_s^T}{\sqrt{d_k}}\right) V_s \tag{2}$$

Where:

- Q_s, K_s, and V_s are the query, key, and value matrices derived from the spatial tokens.
- d_k is the dimension of the keys.

DaViT efficiently collects both large global context and specific local information by switching between these two methods without requiring undue processing overhead.

Computational Efficiency

DaViT addresses the computational inefficiencies typical in standard self-attention, which often exhibits quadratic complexity:

$$O(N^2) \tag{3}$$

Where N is the number of tokens. By grouping feature channels and performing attention within each group, DaViT reduces complexity to a linear scale represented as:

$$O(S + C) \tag{4}$$

Where S and C denote the spatial and channel dimensions, respectively. This design allows channel-wise attention to efficiently capture global context. Additionally, DaViT's dual-attention mechanism combines global representations with fine-grained local detail through spatial window attention, enhancing spatial interactions.[4]

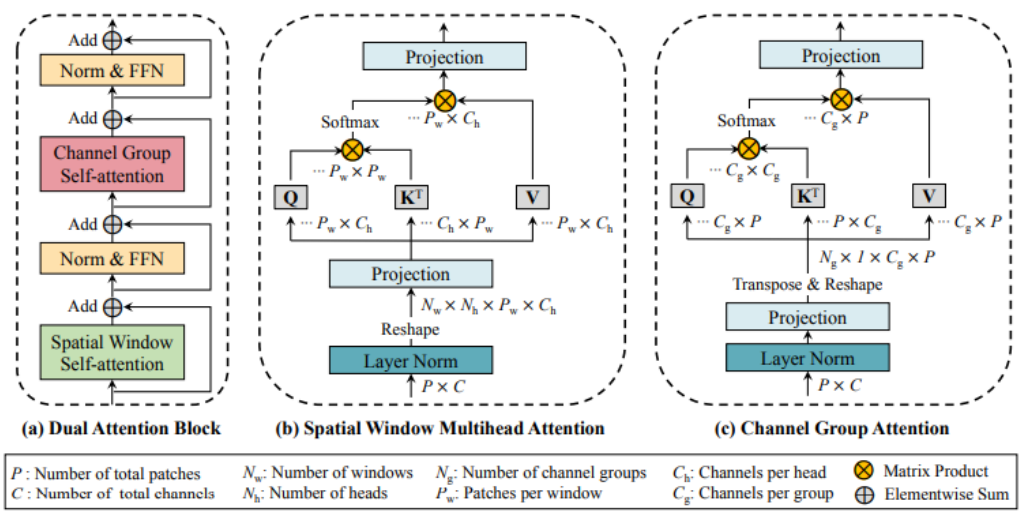

Figure 1: Block diagram of the developed DaViT pipeline. Adapted from [4].

3 Results

3.1 Achieved results on the validation dataset

A number of important metrics, such as the precision-recall curve, ROC curve, and per-class precision, recall, and F1 score, were used to assess the DaViT model's performance on the validation dataset.[8]. These metrics assess the model's overall efficacy across a range of abnormalities and offer insight into how well it can differentiate between classes.

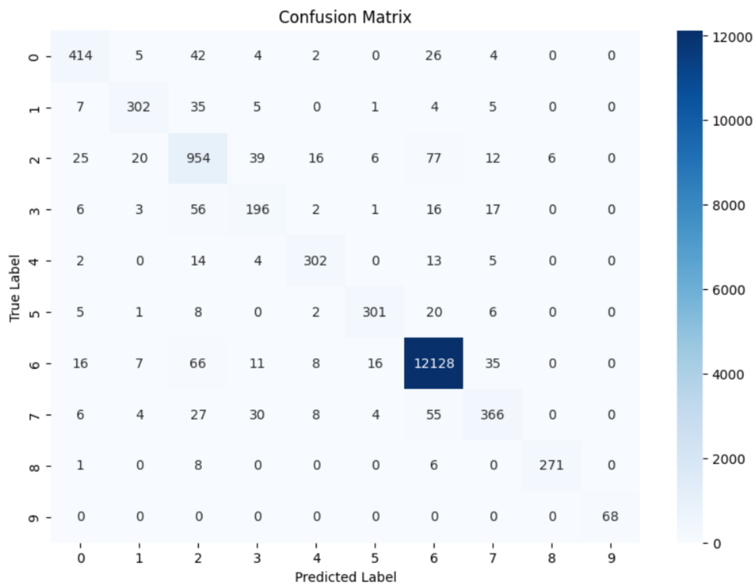

Figure 2: *Confusion Matrix for the DaViT Model on the Validation Dataset.* This matrix shows true versus predicted classifications for each class, with higher diagonal values indicating accurate predictions and off-diagonal values representing misclassifications.

In our experiments, the DaViT model outperformed traditional methods across key metrics. The detailed performance results are shown in Table 1.

Metric	Vit	mViT	DaViT
Accuracy	91.69%	94.40%	94.76%
Precision	80.222%	80.85%	94.83%
Recall	74.921%	87.11%	94.85%
F1-score	76.943%	83.86%	94.83%

Table 1: Performance comparison of models on key metrics.

4 Discussion

We used a custom Convolutional Neural Network (CNN) model to extract features from RGB photos and achieve an accuracy of approximately 85%. The CNN architecture consists of three convolutional layers meant to capture key patterns in the data, followed by a fully connected layer with dropout to reduce overfitting. While the model produced encouraging results, we recognised its limitations in dealing with increasingly complicated patterns, which prompted our investigation of advanced architectures.

To boost our CNN findings, we used the ResNet architecture [7]. ResNet's unique residual connections make deeper networks possible by addressing the vanishing gradient problem, theoretically improving model performance. Despite these advantages, we discovered that ResNet did not significantly beat our initial CNN model. This result demonstrated that simply increasing model complexity was insufficient to provide the expected performance gains.

This realization prompted us to investigate the Vision Transformer (ViT) model [2], which uses self-attention mechanisms to better capture relationships between visual components. The ViT model demonstrated significant performance gains, with an accuracy of 91.69%, precision of 80.22%, recall of 74.92%, and F1-score of 76.94%. Encouraged by these findings, we explored the mViT[6] and DaViT[4] models, which produced even better metrics: mViT achieved 94.40% accuracy with 80.85% precision, 87.11% recall, and an F1-score of 83.86%, whereas DaViT achieved 94.87% accuracy, 94.82% precision, 94.76% recall, and an F1-score of 94.77%.

The resilience of our models is further supported by ROC curve analysis, which shows their overall efficacy. The Area Under the Curve (AUC) values for most classes—including Bleeding, Normal, Ulcer, and Worms—are close to 1.00, showing high discrimination ability. However, the AUC for Erythema is slightly lower at 0.98, indicating that there is opportunity for improvement in separating this category from others.

5 Conclusion

In this study, we addressed the task of multi-class anomaly detection in Video Capsule Endoscopy (VCE) using a variety of deep learning architectures, culminating in the Dual Attention Vision Transformer (DaViT) model. Our thorough investigation found that, while initial efforts with custom CNNs generated an accuracy of around 85%, the addition of advanced architectures such as ResNet[7] and Vision Transformers[2] greatly improved our performance metrics. Notably, the DaViT model produced excellent results, with an accuracy of 94.87%, precision of 94.82%, recall of 94.76%, and an F1-score of 94.77% [8], proving its resilience in differentiating distinct gastrointestinal anomalies.

DaViT's dual attention mechanism[4], which adeptly integrates both spatial and channel-wise attention, has proven useful in capturing the complicated correlations found in video data. The model's computational efficiency, reduced complexity, and capacity to extract hierarchical information have set new standards in the field, implying that it can greatly speed up the diagnostic process in clinical situations.

While our findings are encouraging, they also point to areas for further investigation, notably in enhancing the model's discrimination capabilities for categories such as Erythema. Our findings pave the way for further study into the use of transformer-based models in medical imaging, potentially leading to more accurate and efficient diagnostic tools. The achievements highlighted herein not only add to the current research, but also demonstrate AI's transformative potential in improving healthcare outcomes.

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 [1].

References

[1] P. Handa, A. Mahbod, F. Schwarzhans, R. Woitek, N. Goel, D. Chhabra, S. Jha, M. Dhir, D. Gunjan, J. Kakarla, B. Raman, "Capsule Vision 2024 Challenge: Multi-Class Abnormality Classification for Video Capsule Endoscopy," Research Center for Medical Image Analysis and Artificial Intelligence, Danube Private University, Krems, Austria, 2024.

[2] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[3] Y. Li, C. Y. Wu, H. Fan, K. Mangalam, B. Xiong, J. Malik, and C. Feichtenhofer, "MViTv2: Improved Multiscale Vision Transformers for Classification and Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4804-4814.

[4] G. Ding, H. Zhang, and X. Wang, "Dual Attention Vision Transformer (DaViT)," arXiv preprint arXiv:2204.03645, 2022.

[5] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," Facebook AI Research (FAIR) and UC Berkeley.

[6] Y. Li, C. Y. Wu, H. Fan, K. Mangalam, B. Xiong, J. Malik, and C. Feichtenhofer, "MViTv2: Improved Multiscale Vision Transformers for Classification and Detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4804-4814.

[7] He, K., Zhang, X., Ren, S., & Sun, J. "Deep residual learning for image recognition" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016 , pp. 770-778.

[8] Vermma, D., Saxena, V., Sharma, D., & Gupta, A. (n.d.). *capsule-commandos*. GitHub repository. `https://github.com/devrishivermaa/capsule-commandos`