Capsule Vision Challenge
Team : Layer Players

Satyarth Singh, Praveen, Prerana Mukherjee

School of Engineering
Jawaharlal Nehru University

Email
satyarth2017@gmail.com
prerana@jnu.ac.in

Abstract

This study presents an approach to classify abnormalities in video capsule endoscopy (VCE) frames using a modified ResNet101 architecture with Squeeze and Excitation (SE) blocks. The aim was to build a generalized, model capable of automatic abnormality detection across ten classes: Angioectasia, Bleeding, Erosion, Erythema, Foreign Body, Lymphangiectasia, Polyp, Ulcer, Worms, and Normal. Our approach involved augmenting and balancing the dataset to address class imbalance, followed by training and evaluating the model. The model achieved a mean AUC of 0.984, mean specificity of 0.990, mean average precision of 0.839, mean sensitivity of 0.759, and a balanced accuracy of 0.780. These results demonstrate the potential of SE-ResNet101 for VCE abnormality detection, with opportunities for further improvement in sensitivity and overall accuracy.

1 Introduction

Video capsule endoscopy (VCE) has become a vital tool for non-invasive visualization of the gastrointestinal (GI) tract. However, the sheer volume of video data generated requires automated solutions for abnormality detection to support clinical decision-making. The Capsule Vision Challenge 2024 provides an opportunity to evaluate AI models that can classify various abnormalities from VCE frames across ten categories. The purpose of this challenge is to foster the development of vendor-neutral, generalized models for clinical deployment. This report outlines our approach, which utilizes an augmented and balanced dataset and a modified ResNet101 model with SE blocks for improved feature representation and classification.

2 Methods

2.1 Data Preprocessing and Augmentation

To address class imbalance in the VCE dataset, we augmented and balanced the dataset to 5,000 samples per class, ensuring sufficient representation of each abnormality class. Standard image preprocessing techniques were applied, including resizing, normalization, and random transformations such as rotations, flips, and color adjustments. This balanced dataset provided a robust foundation for training.

2.2 Model Architecture

In our approach, we adopted the ResNet101 architecture, a well-established deep learning model known for its effectiveness in image classification tasks, particularly in complex datasets. To further enhance its performance, we integrated Squeeze and Excitation (SE) blocks within each residual block of the ResNet101 architecture. The inclusion of SE blocks is a pivotal design choice that addresses a common limitation in traditional convolutional neural networks: the inability to adaptively recalibrate feature responses based on their importance.

The Squeeze and Excitation blocks function by performing a two-step process: first, they ”squeeze” global information by aggregating feature maps across spatial dimensions, thereby capturing the channel-wise dependencies. This process generates a compact representation of the feature distribution, allowing the model to discern which features are crucial for the task at hand. Next, in the ”excitation” phase, these features are recalibrated through a learned scaling mechanism. This mechanism emphasizes the most informative features while suppressing less relevant ones, ultimately enabling the model to focus on critical patterns that contribute to classification decisions.

By incorporating the SE blocks into our enhanced ResNet101-SE architecture, we aim to improve the model’s ability to capture intricate and nuanced patterns within the video capsule endoscopy (VCE) frames. These frames often contain subtle abnormality features—such as slight color variations, texture anomalies, or irregular shapes—that can be challenging to differentiate from normal variations in the tissue. The dynamic recalibration of features provided by the SE blocks is particularly beneficial in this context, as it allows the model to adaptively enhance the representation of these subtle cues, leading to improved classification accuracy.

2.3 Training and Validation

The model was trained on the augmented dataset using cross-entropy loss and an Adam optimizer with a learning rate scheduler to optimize performance. Hyperparameters were fine-tuned based on early experiments, focusing on maximizing AUC and balanced accuracy. Validation was conducted on the validation dataset provided by the organisers, with metrics including mean AUC, specificity, sensitivity, mean average precision (mAP), and F1 score.

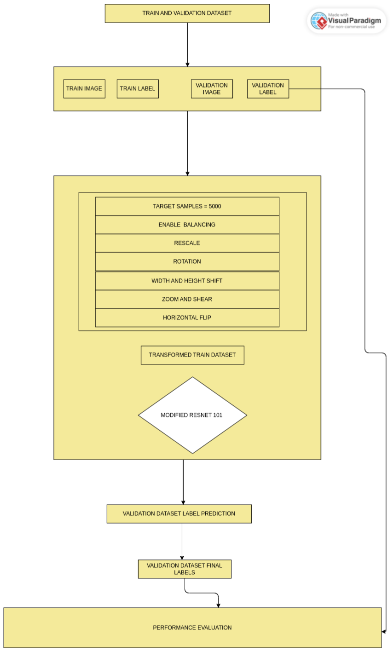

Figure 1: Block diagram of the developed pipeline.

3 Results

Table 1: Achieved results on the validation dataset

Method	Avg. AUC	Avg. Specificity	Avg. Sensitivity	Avg. F1-score	Avg. Precision
ResNet101 with Squeeze-Excitation (Solution Model)	**0.9843**	**0.9896**	**0.7593**	**0.7578**	**0.8388**

Table 2: Performance Comparison of Different Methods

Method	Avg. AUC	Avg. Specificity	Avg. Sensitivity	Avg. F1-score	Avg. Precision
VGG16 (baseline)	0.9161	0.9697	0.5430	0.4844	0.5246
SVM (baseline)	0.94	0.813	0.408	0.487	0.833
ResNet50 (baseline)	0.871	0.814	0.32	0.373	0.60
Custom CNN (baseline)	0.4975	0.898	0.097	0.093	0.10
ResNet101 with Squeeze-Excitation (Solution Model)	**0.9843**	**0.9896**	**0.7593**	**0.7578**	**0.8388**

4 Discussion

The high specificity and mean AUC achieved demonstrate that the ResNet101-SE model is effective in identifying abnormal frames with a low rate of false positives, which is critical for reducing unnecessary follow-ups in a clinical setting. However, the model's sensitivity suggests variability in accurately detecting all abnormal classes, likely due to the subtle nature of certain abnormalities in VCE frames.

5 Conclusion

This study presents an SE-ResNet101-based approach for VCE abnormality classification, achieving high specificity and mean AUC, with competitive sensitivity and F1 scores. Our approach demonstrates that an augmented and balanced dataset combined with SE blocks in a deep learning model can support effective classification of VCE frames into clinically relevant classes. This work contributes to the field by highlighting the feasibility of vendor-independent AI-based models for VCE interpretation and offers insights for further improvements toward clinical deployment.

6 Acknowledgments

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

References

- 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," July 2024.
  Available: `https://figshare.com/articles/dataset/Training_and_Validation_Dataset_of_Capsule_Vision_2024_Challenge/26403469`, doi: 10.6084/m9.figshare.26403469.v1, Figshare.
- 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.