Interpretability-based Robustness Analysis of Medical Image Segmentation Models

Pierre Treyer

Supervisor(s): Prof. Dr. Mauricio Reyes, MSc Zixin Shu
Institution(s): University of Bern, ARTORG Center for Biomedical Engineering Research,
Examiners: Prof. Dr. Mauricio Reyes and Zixin Shu

Introduction

Despite the rapid development and proliferation of novel AI methodologies, many errors produced by AI systems are challenging to interpret, as it frequently function as opaque "black boxes". While previous studies have made significant contributions to improving model transparency, they largely concentrate on the final layers of the model, which may limit the understanding of errors occurring earlier in the decision-making process. In contrast, this thesis aims to explore saliency maps across various layers of the network and to investigate the mapping between failure patterns or so-called failure modes and Magnetic Resonance Imaging (MRI) sequence-specific image perturbations. The research intended to develop a mechanism for flagging incorrect AI results, thereby providing a new approach to quality control.

Materials and Methods

Five specific perturbations, each with five levels of severity, were applied to images from four datasets of the MedMNIST collection. Saliency maps were generated across multiple layers for both the original and perturbed images using the GradCAM method. Differences in saliency maps were quantified using the RMSE metric. The relationship between model interpretability and performance was then analyzed by plotting changes in AUC against RMSE.

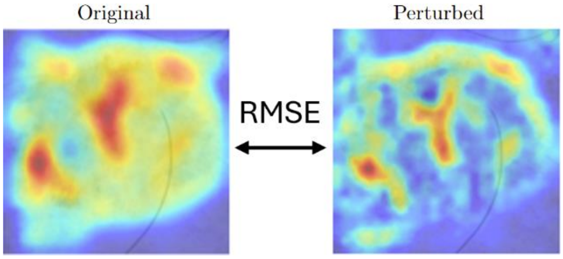

Fig. 1 RMSE between the saliency maps of the original and perturbed image of the 2D DermaMNISTt dataset.

Results

A common trend observed across all datasets was the increase in RMSE values as layers deepened, with significant peaks in the final layers. Additionally, contrast perturbations were particularly disruptive, causing greater fluctuations and higher RMSE values across multiple datasets.

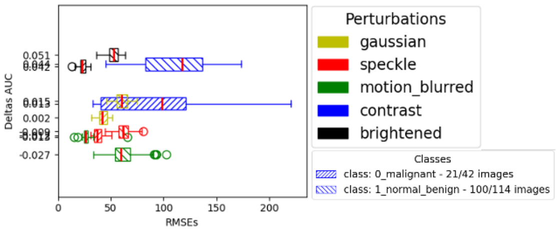

Fig. 2 Correlation between Delta AUC and RMSE for the Breastmnist dataset under perturbations at severity level 3.

A few subtle trends and patterns emerged regarding the model's response to perturbations across every layers. A consistent observation was the linear relationship between delta AUC and RMSE values. Specifically, as severity levels increased, particularly for contrast and brightened perturbations, delta AUC values tended to rise while RMSE values declined, indicating a deterioration in model performance.

Discussion

The consistent presence of error patterns across all network layers, rather than just the final layer, highlights the importance of a comprehensive layer-wise analysis in understanding and improving model behavior. Although the correlations between the performance and RMSE values were limited, those that were found suggest a possible link, particularly under specific perturbations such as contrast and brightness. This specificity suggests that while it is possible to flag incorrect AI results using saliency map discrepancies, this approach may be most effective when tailored to specific types of data and perturbations. More work should be dedicated to developing a reliable mechanism for flagging incorrect AI results.

References

Chattopadhay A., Sarkar A., Howlader P., Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks, IEEE Winter Conference on Applications of Computer Vision, 839-847, 2018.

Acknowledgements

I would like to thank my supervising professor, Mauricio Reyes, and the entire Medical Image Analysis team for their guidance, warm welcome, and support throughout this project.