Project Proposal : Quantitative Auditing of Vision Transformer Explainability in Medical Diagnostics

Members of the group : Nour AFFES, Yunhao ZHOU

Problem

While Vision Transformers have achieved state-of-the-art performance in medical image classification, their clinical deployment is stalled by the "black box" problem. Current Explainable AI evaluations in medical imaging predominantly rely on qualitative visual inspection of saliency maps (Do these red spots look like lungs?). This approach is subjective and prone to confirmation bias. There is a need to rigorously audit whether these explanations faithfully reflect the model's decision-making logic or merely rely on spurious correlations (like hospital tags). This project addresses the requirement for trustworthy AI in healthcare by establishing a quantitative benchmark for model interpretability.

Building on recent critiques of qualitative XAI (Barekatain & Glocker, 2025; Hedström et al., 2023), this project will shift focus from generating explanations to evaluating them.

Dataset : We will use the **Kaggle Chest X-Ray Images (Pneumonia)** dataset (5,863 images), a standard benchmark for binary classification (Normal or Pneumonia).

Methodology

1. **Model Training :** Fine-tune two baseline architectures, a CNN (**ResNet-50**) and a Vision Transformer (**ViT-Base-16**) to achieve comparable accuracy (>90%).
2. **Explanation Generation :** Apply post-hoc importance attribution methods including **Grad-CAM** and **Integrated Gradients** using the Captum library.
3. **Quantitative Auditing :** Use the **Quantus** evaluation framework to measure explainability metrics, specifically **Faithfulness** (for example the pixel flipping impact on confidence) and **Robustness** (stability against Gaussian noise).

Evaluation Strategy

- **Quantitative :** The primary deliverable is a statistical comparison of **Faithfulness Correlation** and **Max-Sensitivity** scores. We hypothesize that while ViTs may have higher accuracy, their explanations may show lower faithfulness compared to CNNs due to the dispersed nature of self-attention.
- **Qualitative :** We will visualize failure cases where heatmaps align with human intuition but fail mathematical faithfulness tests (and vice versa), to highlight the danger of trusting visuals alone.

Conclusion

Unlike other projects that aim to maximize classification accuracy, our work provides a Meta-Evaluation of the safety features of medical AI. By applying the Quantus framework to the specific domain of pneumonia detection, we aim to expose potential reliability gaps in ViT explanations that are often overlooked in standard performance metrics.

Barekatain, L., & Glocker, B. (2025). *Evaluating the Explainability of Vision Transformers in Medical Imaging* (No. arXiv:2510.12021). arXiv. https://doi.org/10.48550/arXiv.2510.12021

Hedström, A., Weber, L., Bareeva, D., Krakowczyk, D., Motzkus, F., Samek, W., Lapuschkin, S., & Höhne, M. M.-C. (2023). *Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond* (No. arXiv:2202.06861). arXiv. https://doi.org/10.48550/arXiv.2202.06861