Explainable AI for Medical Image Segmentation: GradCAM on Pre-trained U-Net

Jean-Baptiste Felici
Student Number: G20250339
Art & Technology in Sogang University
jefe61830@eleve.isep.fr

Abstract

Deep learning models like U-Net excel in medical image segmentation but lack interpretability, critical for clinical trust. This project applies GradCAM, a gradient-based XAI method, to a pre-trained U-Net to generate heatmaps highlighting influential regions in segmentation tasks (e.g., tumor detection). We propose an interactive visualization tool using Streamlit for user-friendly exploration. Expected outcomes include quantitative evaluation via Deletion/Insertion metrics and improved understanding of U-Net's decision-making.

1. Introduction

Medical image segmentation using deep learning (e.g., U-Net [1]) has transformed diagnostics, yet its black-box nature raises concerns in clinical settings. Explainable AI (XAI) methods, such as GradCAM [2], offer visual insights into model decisions. This project extends GradCAM to segmentation, leveraging course insights on gradient-based attribution, to enhance transparency in pre-trained models.

2. Approach

We will use a pre-trained U-Net (e.g., from Hugging Face or Kaggle datasets like BraTS) for segmenting medical images (e.g., MRI scans). GradCAM will compute gradients from decoder layers, producing class-discriminative heatmaps for each segmented class (e.g., tumor vs. background). An interactive Streamlit dashboard will allow users to upload images, view segmentations, and toggle GradCAM overlays, inspired by course visualizations.

2.1. Implementation Plan

1. Load a pre-trained U-Net via PyTorch. 2. Implement LayerGradCam from Captum on U-Net's decoder layers. 3. Generate heatmaps and build a Streamlit app. 4. Evaluate with Deletion/Insertion Game and ROAR metrics on a small dataset (e.g., 100 MRI slices).

2.2. Expected Outcomes

- Heatmaps revealing U-Net's focus (e.g., tumor edges vs. artifacts). - Quantitative scores (e.g., low Deletion AUC for accurate attributions). - A demo of the interactive tool, enhancing clinical interpretability.

References

[1] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *arXiv preprint arXiv:1505.04597*, 2015. 1

[2] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization, 2016. 1