Decoding Neural Netowrks - Report

Saranath P

Table of contents

1 Project Title

Decoding Neural Networks: Explainability of Brain Tumor Classifications

2 Brief Description

In today's medical world, it is crucial to have a clear understanding of why machine learning models make certain predictions, especially in sensitive areas like brain tumor diagnosis. This project focuses on using explainability techniques such as **GradCAM**, **Sparse AutoEncoders**, and **Saliency Maps** to analyze why neural networks classify brain tumors into specific categories. By applying these techniques to a DenseNet model, we aim to demonstrate whether the neural networks are making meaningful and justifiable predictions. There are 4 classes in dataset.

- Glioma
- Meningioma
- Notumor
- Pitutary

3 Technology Stack

- **Languages**: Python
- **Frameworks**: PyTorch
- **Explainability Techniques**: GradCAM, Sparse AutoEncoders, Saliency Maps
- **Models**: DenseNet169, Vision Transformers (ViT)
- **Tools**: Custom-built explainability modules and PyTorch inbuilt functions for model customization

4 Challenges and Solutions

- **Challenge 1**: Fine-tuning both DenseNet and Vision Transformer (ViT) models was particularly challenging due to the complexity and size of the datasets.
  - **Solution**: Optimized the models by adjusting hyperparameters and using smaller subsets of the dataset to achieve better training results.
- **Challenge 2**: Understanding how GradCAM works and modifying it to fit the current task of brain tumor classification required significant effort.

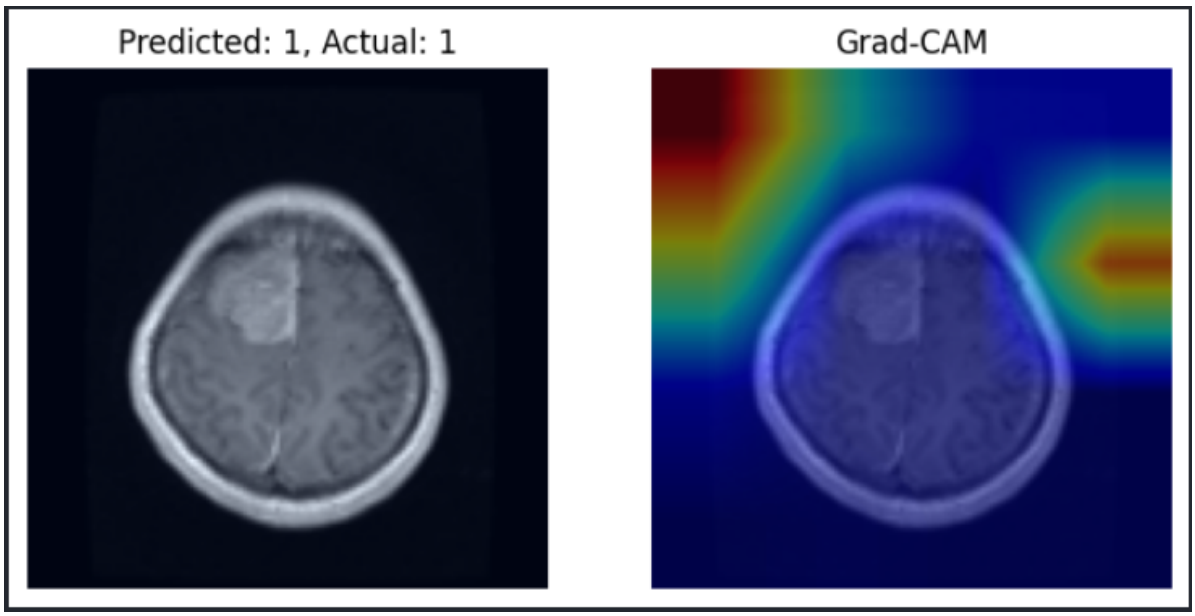
– **Solution**: We customized GradCAM from scratch to handle the specific requirements of brain tumor classification and integrated Sparse AutoEncoders to enhance the visualization of the important regions in brain images.

5 Team Size

This project was completed by a team of **two members**. My specific contributions included developing the custom GradCAM and Sparse AutoEncoders from scratch and fine-tuning the DenseNet169 model.

6 Project Outcome

- **Accuracy**: The DenseNet169 model achieved an accuracy of **0.93** on the test set.
- **Key Findings**: Some model predictions were accurate but lacked meaningful explanations. For example, some pixels outside the brain were deemed important by the model, highlighting potential issues in how neural networks learn and generalize.
- **Impact**: These findings demonstrate the importance of explainability in medical applications and can help guide further research into refining explainability techniques for deep learning models in healthcare.



Here **Label 1** refers meningioma. There are other several images like this where even though the prediction made was right, the features it took to make the prediction was not even inside the brain image.

7 Links

- **Project Repository**: Link to GitHub Repo
- **Demo/Documentation**: Link to Documentation or Demo