Explainable AI for Brain Tumor MRI Segmentation: Grad-CAM on a Pre-trained U-Net

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

Abstract

Deep learning models like U-Net achieve strong performance on medical image segmentation but remain difficult to interpret for clinicians. This project applies Grad-CAM, a gradient-based explainable AI method, to a U-Net with a ResNet34 encoder pre-trained on ImageNet, fine-tuned on a public brain tumor MRI dataset with pixel-wise tumor masks. The model produces brain tumor segmentations, and Grad-CAM is adapted from a segmentation setting to a scalar output through a simple wrapper, yielding heatmaps that highlight spatial regions influential for the tumor prediction. Qualitative visualizations on MRI slices and a causal Deletion/Insertion evaluation illustrate how much the prediction score depends on Grad-CAM-identified pixels, providing insight into whether the model focuses on tumor regions or spurious background.

1. Introduction

Medical image segmentation has become a central application of deep learning in clinical workflows, especially for tasks like brain tumor delineation on MRI scans using architectures such as U-Net. [2] Despite their strong performance, segmentation models are often perceived as black boxes, which limits trust and adoption in high-stakes environments such as radiology and neurosurgery. Explainable AI (XAI) methods, in particular gradient-based attribution techniques like Grad-CAM, offer a way to visualize which regions of an image contribute most to a model's prediction by generating class-discriminative heatmaps. [3, 4]

Figure 1 shows examples of meningioma MRI slices from the dataset with ground-truth tumor masks overlaid, illustrating the type of lesions the model aims to segment.

This project focuses on making a U-Net segmentation model for brain tumor MRI more interpretable by adapting Grad-CAM to the pixel-wise segmentation setting. Instead of training a model from scratch, a U-Net with a ResNet34 encoder pre-trained on ImageNet is fine-tuned on a public brain tumor dataset containing T1-weighted MRI slices with expert-annotated tumor masks. The goal is to show, for each prediction, which spatial regions drive the network's decision and to study the causal impact of these regions through a Deletion/Insertion evaluation game. [1]

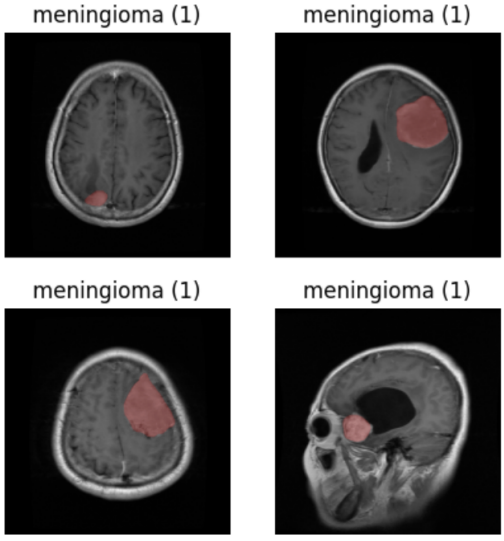

Figure 1. Examples of meningioma MRI slices with ground-truth tumor masks overlaid.

2. Related Work

U-Net for medical image segmentation. U-Net and its variants have become standard for biomedical segmentation, thanks to their encoder–decoder architecture with skip connections, which allows precise localization from relatively few training samples. [2] Numerous works have applied U-Net to brain tumor segmentation in MRI, but most models remain opaque in terms of how they localize lesions.

Gradient-based explanations and Grad-CAM. Grad-CAM produces visual explanations by backpropagating gradients to intermediate feature maps and aggregating them into coarse heatmaps that highlight important regions

for a target class. [3] Originally developed for image classification and related to earlier class activation mapping approaches, [4] Grad-CAM has been adapted to detection and segmentation tasks by treating the model as a feature extractor and carefully choosing the target layer, typically in deeper decoder or classification layers. In medical imaging, Grad-CAM has been used to analyze whether models attend to clinically meaningful structures or to artifacts such as text, borders, or acquisition patterns.

Causal evaluation of explanations. Beyond visual inspection, recent work emphasizes causal evaluation of saliency maps using perturbation games such as Deletion and Insertion, where pixels ranked as important are progressively removed or added and the impact on the model score is measured. [1] A good explanation should lead to a rapid drop in prediction confidence under deletion and a rapid increase under insertion, which can be quantified via the area under the curve (AUC) of the score vs. perturbation steps. In this project, a simplified Deletion/Insertion game is implemented for the segmentation setting to assess whether Grad-CAM truly captures pixels that drive the tumor prediction. [1]

3. Methods

3.1. Model: Pre-trained U-Net with ResNet34 Encoder

The core segmentation model is a U-Net implemented with the `segmentation_models_pytorch` library, using a ResNet34 encoder initialized with ImageNet weights and a single-channel output for binary tumor segmentation. The network takes 2D MRI slices as input resized to 224×224 and outputs a logits map of the same spatial resolution, which is transformed into tumor probability maps via a sigmoid activation during inference. Training uses the Adam optimizer with a learning rate of 1×10^{-4} and a binary cross-entropy loss with logits (BCEWithLogitsLoss), optimized for a small batch size due to GPU memory constraints.

To leverage the pre-trained encoder, grayscale MRIs are converted to pseudo-RGB by duplicating the single channel into three channels, which matches the ResNet34 input format. The model is trained for a small number of epochs mainly to demonstrate the end-to-end pipeline rather than to reach state-of-the-art segmentation accuracy, and the resulting weights are saved for later evaluation and visualization.

3.2. Brain Tumor MRI Dataset and Preprocessing

The dataset is the public brain tumor MRI collection *brainTumorDataPublic_1–766*, distributed as MATLAB `.mat` files with a `cjdata` structure containing label, image, and tumor mask for each patient slice. Each sample includes a T1-weighted MRI slice, a binary ground-truth tumor mask, and an integer label indicating one of three tumor types: meningioma, glioma, or pituitary tumor.

All `.mat` files are converted to NumPy arrays `images.npy`, `masks.npy`, and `labels.npy`, after resizing the image and mask to a fixed resolution for storage and later to 224×224 for training. Images are normalized by their maximum pixel value to obtain intensities in $[0, 1]$, and masks are stored as boolean or $\{0, 1\}$ arrays to represent tumor vs. background.

3.3. Segmentation Training Pipeline

A custom PyTorch `Dataset` wraps the MRI images and masks, performing resizing, normalization, grayscale-to-RGB conversion, and tensor conversion on the fly. A `DataLoader` iterates over the dataset with small batches and shuffling, while using a limited number of worker processes to balance speed and resource usage.

During training, the model outputs logits maps for each batch, which are compared to the ground-truth masks using BCEWithLogitsLoss. The training loop tracks the per-batch loss with `tqdm` progress bars and prints the average loss per epoch; after training, the model is switched to evaluation mode and its weights are saved to `unet_trained.pth`.

3.4. Grad-CAM for Segmentation via a Wrapper

Grad-CAM is designed for scalar outputs, whereas U-Net produces dense segmentation maps. [3] To bridge this gap, a `SegmentationWrapper` module is defined around the U-Net: it forwards the input through the original network and then averages the output logits spatially over height and width, yielding a single scalar per image. This scalar can be interpreted as an overall "tumor intensity" score, which serves as the target for Grad-CAM attribution.

LayerGradCam from the Captum library is instantiated with the wrapped model and a target layer chosen in the decoder head, so that gradients flow through high-level spatial features closely related to the output mask. For a given input batch, LayerGradCam computes attribution maps in the target layer's resolution, which are then upsampled via bilinear interpolation to match the input image resolution and normalized to form Grad-CAM heatmaps.

Figure 2 shows an early experiment where Grad-CAM is applied before fine-tuning on the medical dataset, revealing noisy and diffuse patterns that motivate domain-specific training.

3.5. Batch Grad-CAM and Conditioning on Ground Truth

To analyze more than one image at a time, a batch Grad-CAM function prepares a batch of MRIs by resizing, normalizing, and replicating the grayscale channel into RGB, then feeds the batch through LayerGradCam. The resulting

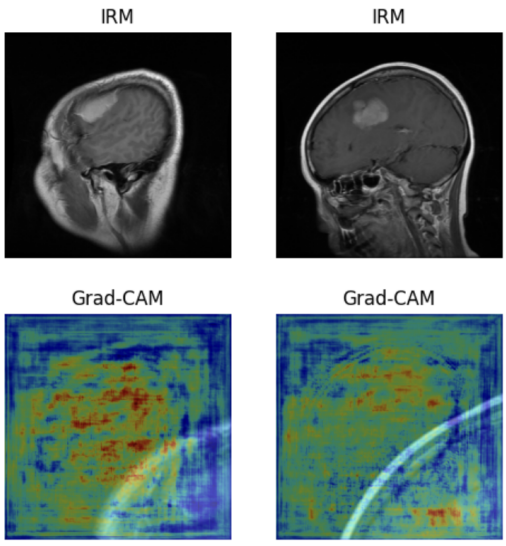

Figure 2. Example MRI slices and Grad-CAM maps before fine-tuning the U-Net on the brain tumor dataset.

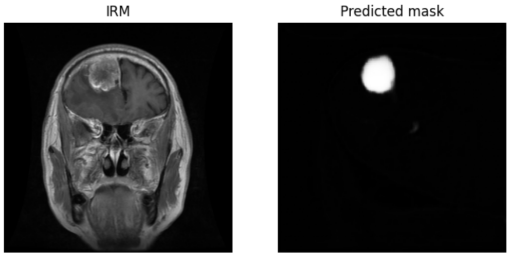

Figure 3. Examples of MRI slices (top) and Grad-CAM $\times$ ground-truth maps (bottom), showing alignment between model attention and tumor regions after training.

tensor of Grad-CAM maps is interpolated back to the input size, and each heatmap is visualized together with its corresponding MRI slice. Additionally, a masking function multiplies each Grad-CAM map by the corresponding ground-truth tumor mask, after resizing if necessary. This produces "Grad-CAM $\times$ GT" maps that highlight only the regions where both the explanation and the annotated tumor overlap, making it easier to visually assess alignment between model attention and the true lesion.

Figure 3 illustrates examples where Grad-CAM mostly lies inside the tumor region and one case where attention is more focal and localized.

3.6. Causal Deletion/Insertion Game

To go beyond qualitative impressions, a causal Deletion/Insertion game is implemented for the segmentation setting. [1] Given a single MRI input and its Grad-CAM map, the heatmap is flattened and pixels are sorted in descending order of importance, defining a ranking from most

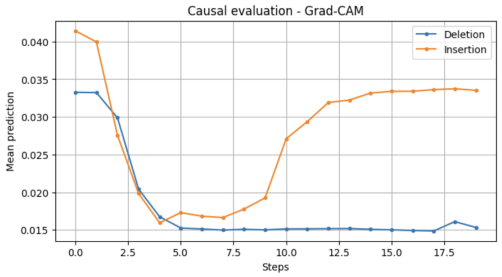

Figure 4. Causal evaluation of Grad-CAM using the Deletion and Insertion game: mean prediction score as a function of perturbation steps.

to least important.

In the Deletion mode, an increasing fraction of the most important pixels is masked out from the input image at each step by multiplying them with zero, while the rest of the image remains unchanged; in the Insertion mode, the image starts from zeros and important pixels are added back progressively. At each step, the modified image is passed through the U-Net and the mean sigmoid output over all pixels is recorded as a global tumor-score. Figure 4 plots the prediction score as a function of perturbation steps for both games.

4. Data Description

4.1. Dataset Structure

The brain tumor dataset used in this project consists of 766 individual MRI slices stored as separate `.mat` files in a `data/` directory, named sequentially from `1.mat` to `766.mat`. Each file contains a `cjdata` MATLAB structure with fields that include the integer class label, the 2D MRI image, and a binary tumor mask of the same resolution. After conversion, three NumPy arrays are stored on disk: `images.npy` for grayscale MRI slices, `masks.npy` for binary tumor masks, and `labels.npy` for the tumor class labels.

4.2. Preprocessing Choices and Limitations

Several practical choices were made to simplify the pipeline. All images are resized to a uniform resolution before training and Grad-CAM computation, which may slightly distort shapes but greatly simplifies batching and visualization. The project uses a relatively small number of training epochs and does not implement a dedicated validation split or hyperparameter search, so the model's segmentation performance is not fully optimized.

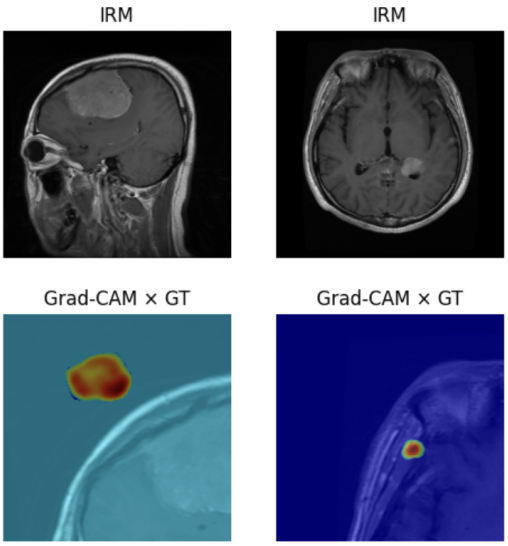

Figure 5. Qualitative example after training: input MRI (left) and predicted tumor mask (right).

5. Experiments and Results

5.1. Qualitative Segmentation Results

After training, the U-Net model is evaluated qualitatively by selecting random MRI slices from the dataset and visualizing both the original MRI and the predicted tumor mask. Figure 5 shows a typical example where the model successfully highlights a large tumor region near the cortex.

The predicted masks demonstrate that the model learns to detect bright tumor regions and can roughly reproduce the shape of the lesions, although some boundaries appear coarse or partially missing due to the limited training budget. These visualizations serve as a sanity check for the model before applying Grad-CAM: if the segmentation were completely random, interpretation of the explanations would be meaningless.

5.2. Grad-CAM Heatmaps on MRI Slices

Grad-CAM heatmaps are generated for both individual images and small batches. [3] For each MRI input, the scalar output of the SegmentationWrapper is used as the target for LayerGradCam applied to the segmentation head of the U-Net, and the resulting heatmap is resized to match the input resolution and overlaid on the MRI.

On many examples, the Grad-CAM maps concentrate on the tumor region or its edges, indicating that the model relies heavily on the lesion for its global tumor score. Conditioning the Grad-CAM maps with the ground-truth mask, as in Fig. 3, highlights the overlap between attention and lesion and visually confirms that a significant fraction of the explanation falls inside the annotated tumor area.

5.3. Causal Evaluation with Deletion and Insertion

The Deletion/Insertion game is run on randomly selected MRI slices to assess the causal impact of Grad-CAM-ranked pixels on the model's global tumor score. [1] For Deletion, as progressively larger sets of high-importance pixels are removed from the input, the mean sigmoid output over the segmentation map tends to decrease, which indicates that Grad-CAM is capturing regions that significantly contribute to the prediction. Similarly, in the Insertion mode, starting from a blank input and gradually inserting the most important pixels leads to an increasing tumor score, supporting the idea that these pixels encode critical information for the model.

Although the implementation mainly focuses on plotting the score versus perturbation steps rather than computing a formal AUC metric, the qualitative shape of the curves in Fig. 4 is consistent with meaningful explanations: sharp decline for Deletion and sharp rise for Insertion.

5.4. Discussion and Limitations

Overall, the experiments show that Grad-CAM can be adapted to a segmentation model via a simple scalar wrapper and still provide informative spatial explanations for tumor segmentation. [3] However, the approach has limitations: the scalar reduction loses information about pixel-wise predictions, the choice of target layer is heuristic, and the causal evaluation is based on global scores rather than clinically relevant metrics such as Dice or IoU. [1]

Despite these constraints, the implemented pipeline—from dataset preparation to training, Grad-CAM visualization, and Deletion/Insertion analysis—provides a concrete, working example of explainable AI for medical image segmentation.

References

[1] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *NeurIPS Workshop on Human-Centric Machine Learning*, 2019. Google Brain. 1, 2, 3, 4

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

[3] 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. In *ICCV*, 2017. 1, 2, 4

[4] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016. 1, 2