IMPROVING ADVERSARIAL ROBUSTNESS OF MEDICAL IMAGE ANALYSIS MODELS VIA BLACK-BOX DEFENSE INTEGRATING IMAGE RECONSTRUCTION.

Tran Nhat Khoa **Ly Nguyen Thuy Linh** **Le Tran Quoc Khanh**

University of Information Technology - National University of Vietnam

Department of Computer Science

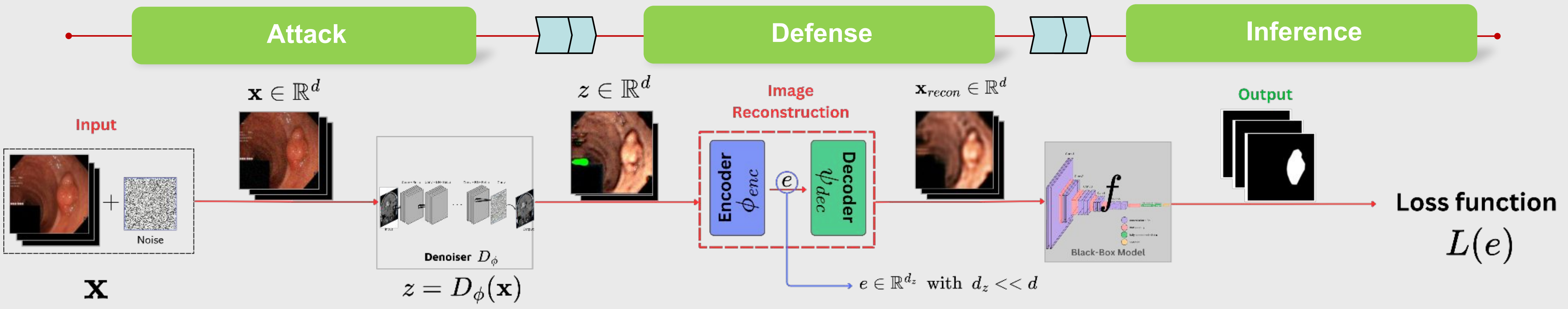
Motivations

Deep neural networks (DNNs), especially convolutional neural networks (CNNs), excel in many computer vision tasks but are vulnerable to adversarial attacks via small input perturbations. Previous methods, like *adversarial training*, require access to the victim model architecture, which is impractical in the medical field due to privacy and security concerns. Zhang et al. proposed a *black-box defense* using *Zeroth-Order Optimization* (ZO), which addresses adversarial attacks while maintaining model confidentiality. ZO approximates model parameters using only input and output but suffers from **high variance** due to the high resolution of medical images. Our research aims to **enhance black-box defense** in medical image analysis models and **resolve ZO's high variance** issue through image reconstruction techniques.

Targets

- Successfully simulate various **adversarial attack algorithms** on medical datasets.
- Protect medical models by applying **black-box defense** on these datasets, integrating **image reconstruction techniques**.
- Provide knowledge, mathematical proofs, and experimental evaluations of black-box defense methods against adversarial attacks on medical image analysis models.

Overview



Description

1. Medical Datasets

For classification, we experiment with the VGG architecture on the MRI Brain Tumor dataset (**A**). For segmentation, we use the U-Net architecture, known for its effectiveness in segmentation tasks, on the Kvasir-SEG dataset (**B**), designed for gastrointestinal image segmentation.

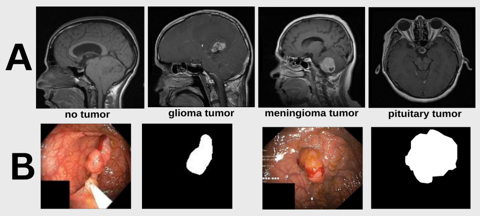


Figure 1 . Medical Datasets

2. Adversarial Attacks

We use popular adversarial attack algorithms to create adversarial examples from original medical datasets. These algorithms include the Fast Gradient-Sign Method (FGSM), Projected Gradient Descent (PGD), and Decoupled Direction and Norm Attack (DDN). The common feature of these attack algorithms is their ability to create perturbations based on gradients, resulting in adversarial examples that are imperceptible to the human eye but can still fool the victim models.

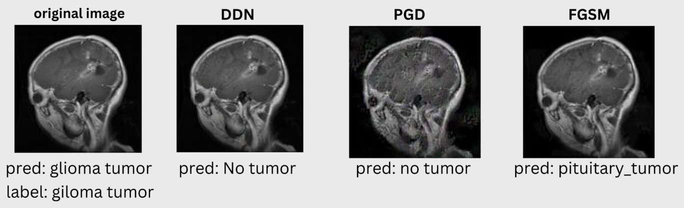


Figure 2 . Adversarial Attacks results on MRI Brain Tumor dataset with Classification task

3. Black-Box Defense

1. Denoised Smoothing (DS)

A denoising-based defense system operates by prepending a denoising neural network (with learnable parameters) to the target model. These learnable parameters can be optimized by minimizing a regularized loss function using any first-order method. However, this requires the gradient of the target model, which is often infeasible with medical models.

$$\mathcal{L}_{DS}(\mathbf{w}) = \mathbb{E}_{\delta\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{I}),\mathbf{x}\sim\mathcal{X}}\Big[\underbrace{\|D_{\mathbf{w}}(\mathbf{x}+\delta)-\mathbf{x}\|^2}_{\text{denoising loss}} + \gamma\underbrace{\mathcal{L}_{CE}(f(D_{\mathbf{w}}(\mathbf{x}+\delta)), f(\mathbf{x}))}_{\text{stability loss}}\Big],$$

Equation 1 . Regularized Loss Function of Denoised Smoothing

The gradient can be approximated via randomized gradient estimation (RGE) using only input and output of model. However, due to the large resolution and high dimensionality of medical images, gradient estimation in RGE suffers from high variance.

$$\hat{\nabla}_{\mathbf{z}} f(\mathbf{z}) = \frac{1}{q}\sum_{i=1}^{q}\left[\frac{d}{\mu}(f(\mathbf{z}+\mu\epsilon_i) - f(\mathbf{z}))\epsilon_i\right],$$

Equation 2 . Randomized Gradient Estimation (RGE)

2. Image Reconstruction

To address the dimensionality problem in medical images, we recommend using **image reconstruction techniques**. Specifically, after being compressed into a lower-dimensional space, the image is reconstructed back into its original space before being fed into the medical image analysis model. This way, we only need to approximate the gradient of the model based on the lower-dimensional vector, thereby solving the initial problem.

Zhang et al. proposed a Black-Box Defense to resolve the high-dimensional problem by placing an AutoEncoder between the denoiser and the target model, called the ZO-AE-DS system. We experimented with ZO-AE-DS on the Kvasir-Seg dataset, a collection of gastrointestinal images. The example result as shown below.

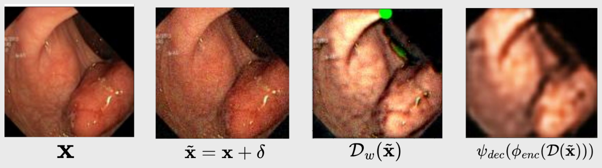

$\mathbf{x}$
$\tilde{\mathbf{x}} = \mathbf{x} + \delta$
$\mathcal{D}_w(\tilde{\mathbf{x}})$
$\psi_{dec}(\phi_{enc}(\mathcal{D}(\tilde{\mathbf{x}})))$

Figure 3 . ZO-AE-DS on Kvasir-Seg dataset

4. Research Plan

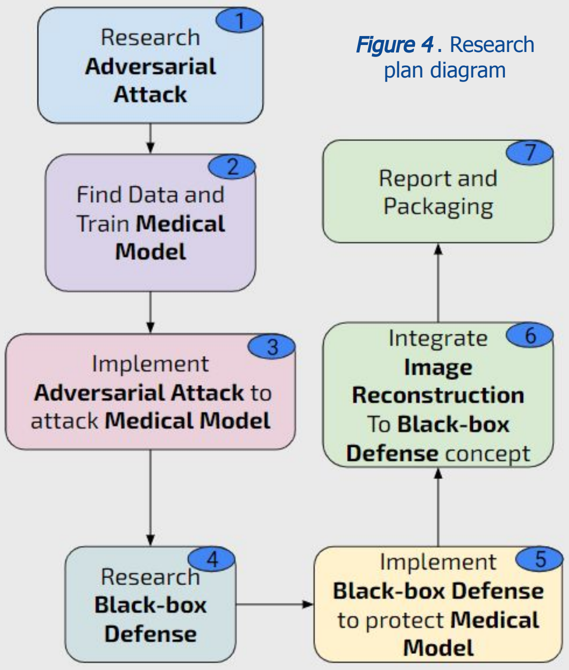

1 Research **Adversarial Attack**
2 Find Data and **Train Medical Model**
3 Implement **Adversarial Attack** to attack **Medical Model**
4 Research **Black-box Defense**
5 Implement **Black-box Defense** to protect **Medical Model**
6 Integrate **Image Reconstruction** To **Black-box Defense** concept
7 Report and Packaging

Figure 4 . Research plan diagram

NII

Tran Nhat Khoa – University of Information Technology
TEL : 0943348069 Email : 22520691@gm.uit.edu.vn