A DE-PSO based attack method for robustness evaluation of Medical Image segmentation

Tushar Sain(187268)[1],Saketh Reddy(187269)[2], and Ajay Yadav (167203)[3]
ts821802@student.nitw.ac.in[1], sr931853@student.nitw.ac.in[2], and yajaykamlesh@student.nitw.ac.in[3]
Department of Computer Science and Engineering, National Institute of Technology, Warangal.

Abstract

Machine learning models such as Image classification and Image segmentation are an effective and reliable tool to assist medical professionals and scientists of various fields with many time-consuming and error-prone medical image analytical tasks. However,recently deep models have been shown to be vulnerable to adversarial attacks,these Adversarial attacks pose significant threat to analytical machine learning models in all fields especially medical. Existing works regarding the robustness of deep learning models are scarce, where most of them focus on the adversarial attacks of medical image classification models. In this project, a hybrid two-phase alternating global optimization algorithm called the DE-PSO method is proposed to generate adversarial examples for medical image segmentation models. The proposed method can achieve better results than gradient-based methods and evolutionary methods such as GA, DE, and PSO which can successfully attack the segmentation model while only perturbing a small fraction of the image pixels with certain noise pixels, demonstrating that the medical image segmentation model is much more susceptible to adversarial examples as compared to medical image classification models.

Introduction

The advancement of Deep Neural Networks and machine learning models in Computer Vision has improved the performance of many computer-aided diagnostic (CAD) systems. Deep learning-based CAD systems are currently being identified as promising tools for supporting doctors and scientists with various time-consuming and error-prone components of their everyday activities. Nonetheless,deep learning frameworks have security flaws: introducing a little, even microscopic amount of adversarial perturbations is enough to change diagnostic results, indicating a significant crisis in the field of medical image analysis. An Adversary can also enter a medical database and utilise this approach to easily invalidate medical data. As a result, the healthy progression of medical research may be hampered.In summary, the Robustness evaluation of any medical segmentation model is deeply necessary

Objective

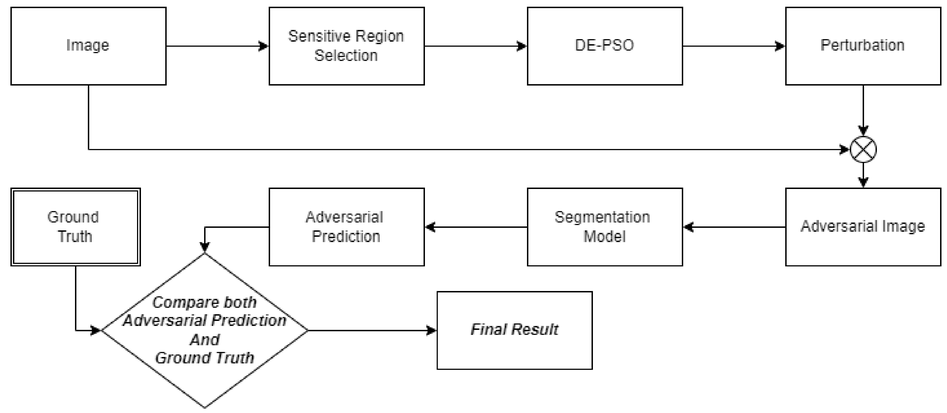
We investigate the robustness evaluation of deep segmentation models in medical image processing in this paper. We show how segmentation results can be successfully altered by changing only a small percentage of image pixels. Specifically, using a two-phase Differential Evolution and Particle Swarm Hybrid assault with multiple iterations, our suggested method can automatically identify the most sensitive regions of input photos. Furthermore, DE-PSOAttack modifies fewer pixels and does not require extra information such as the segmentation network's structures and weights to complete the attack.

Proposed Solution

Our technique is divided into three basic modules, as shown in Figure 1: sensitive region (SR) selection, local optimum detection, and global optimum detection . In the sensitive region selection, we look for the area that corresponds to the foreground of the original image's segmentation mask. Following that, a differential evolution is employed to generate potentially adversarial perturbations. Finally, we check to see if the optimal solution was obtained; if not, a particle swarm is generated to further optimise the perturbations and determine whether more evolution stages are required. The latter two procedures are repeated until the attack is successful or the maximum number of iterations is reached.

- First we obtain the sensitive region information from the foreground of ground truth of the corresponding original image.
- Second we use the DE-PSO attack method to generate perturbation points within the sensitive region of the image.
- Third we add those perturbation points to the original image.
- Fourth we pass the adversarial image into the segmentation model to get the adversarial prediction.
- Finally the Original ground truth and Adversarial mask are compared to get the final results.

Model Overview



DE-PSO Attack method

- **Initialisation** - generated initial population based on sensitive region selection and uniform distribution.
- **DE** -
  - **variation** -the randomly selected individual vector plus the weighted difference vector is used as the variant individual.
  - **crossover** -the crossover operation uses the method of randomly reorganizing the dimensional components of the variant individual and the target individual to generate crossover individuals,with the purpose of increasing the diversity of the population.
  - **selection** -the selection operation is carried out between the target individual and the crossed individual, and through the competition rule of survival of the fittest, select the more adaptable individual to enter the offspring to continue to multiply.
  - **fitness function evalutation** - In this method first we are going to generate adversarial mask using current perturbation points and then feed this image into our segmentation network and generate adversarial mask and compute the fitness value by comparing it with our ground truth and return the fitness value.
- **Comparing fitness values** - If the above fitness value computed using DE is not better than our current best solution then we will use PSO else we will update the current best fitness value and continue with next iteration.
- **PSO** -
  - **velocity update**- over the iterations in the search space, the speed of each particle is stochastically accelerated towards its previous best position (personal best) and towards the best solution of the group (global best).
  - **position update** - Concretely, at each iteration, each particle is updated according to its velocity.
  - **fitness function evaluation** - Again, in this method first we are going to generate adversarial mask using current perturbation points and then feed this image into our segmentation network and generate adversarial mask and compute the fitness value by comparing it with our ground truth and return the fitness value.
- **Updation Best Fitness** - The above fitness value computed using PSO ,we will update the current best fitness value after comparing it with above computed fitness value and continue with next iteration.

Experimental Results

We have varied the number of perturbation points as well as the number of iterations and we have concluded that in both case our proposed DEPSO attack algorithm performs better than the PSOattack(implemented by us in previous semester) algorithm and also results do not vary that much in case of changing number of perturbation points and the number of iterations but results may vary in good way if we increase the values of number of iterations and number of perturbation points because right now we are not able to do that due to lack of computational power.

Average SSIM(in %) vs Number of perturbation points chart :

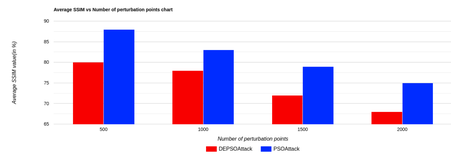


Figure 1:DEPSOAttack vs PSOAttack chart-1.

Average SSIM(in %) vs Number of iterations chart :

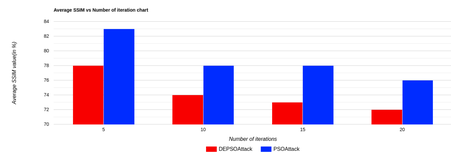


Figure 2:DEPSOAttack vs PSOAttack chart-2.

References

[1] DEAttack: A differential evolution based attack method for the robustness evaluation of medical image segmentation, 2021.

[2] They Might NOT Be Giants Crafting Black-Box Adversarial Examples Using Particle Swarm Optimization, 2020.

[3] Hybrid Differential Evolution - Particle Swarm Optimization Algorithm for Solving Global Optimization Problems, 2008.

GitHub Link

- https://github.com/saketh2341/Project