IMPACT OF ADVERSARIAL EXAMPLES ON DEEP LEARNING MODELS FOR BIOMEDICAL IMAGE SEGMENTATION

Utku Ozbulak[1,3], Arnout Van Messem[2,3], Wesley De Neve[1,3]

[1] Department of Electronics and Information Systems, Ghent University, Belgium
[2] Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Belgium
[3] Center for Biotech Data Science,Ghent University Global Campus, South Korea

Abstract

Deep learning models, which are increasingly being used in the field of medical image analysis, come with a major security risk, namely, their vulnerability to adversarial examples. Given that a large portion of medical imaging problems are effectively segmentation problems, we analyze the impact of adversarial examples on deep learning models for biomedical image segmentation. We expose the vulnerability of these models to adversarial examples by proposing a novel algorithm, namely, the Adaptive Segmentation Mask Attack (ASMA). This algorithm makes it possible to craft targeted adversarial examples that come with high Intersection-over-Union rates between the target adversarial mask and the prediction, as well as with perturbation that is mostly invisible to the bare eye.

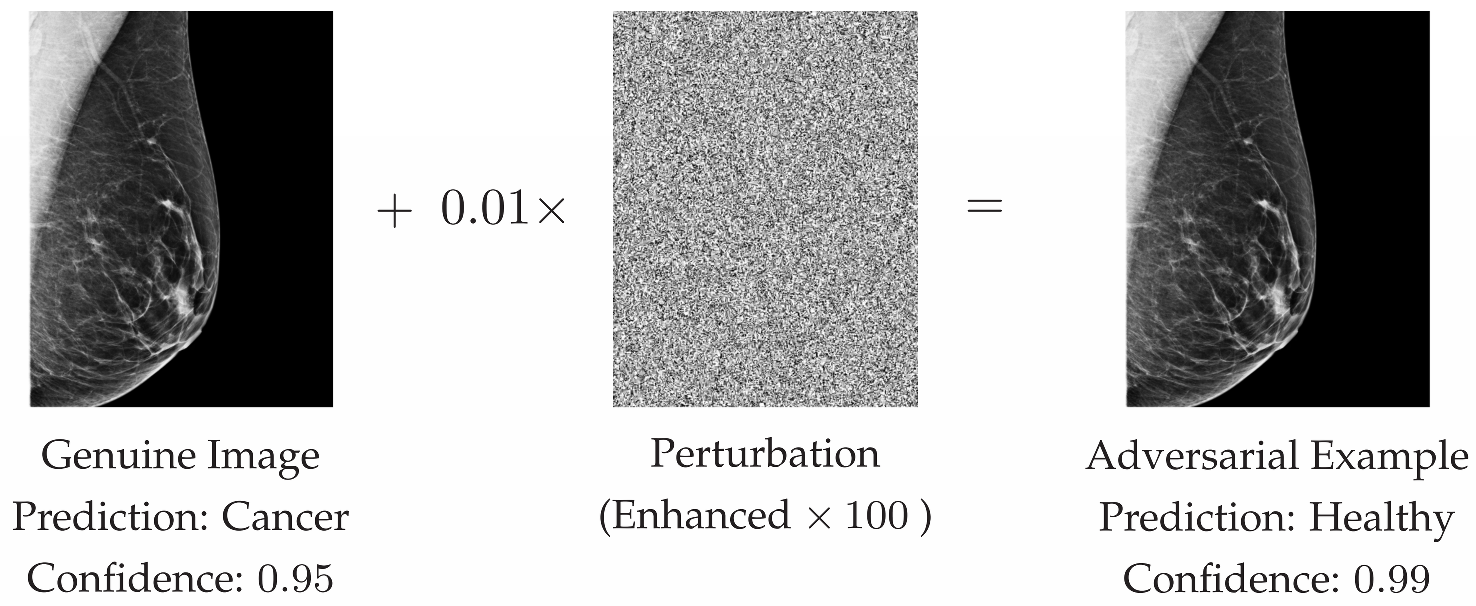
Motivation

Given that (1) labor expenses (i.e., salaries of nurses, doctors, and other relevant personnel) are a key driver of high costs in the medical field and (2) that increasingly super-human results are obtained by machine learning systems, an ongoing discussion is to replace or augment manual labor with *automation* for a number of medical diagnosis tasks [1]. However, a recent development called *adversarial examples* showed that deep learning models are vulnerable to gradient-based attacks [2]. This vulnerability, which is considered a major security flaw, for instance enables the creation of fraud schemes (e.g., for insurance claims) when deep learning models are carrying out clinical tasks [1].

The above observations motivate our effort to better understand the impact of adversarial examples on deep learning approaches towards biomedical image segmentation, so to facilitate the *secure* deployment of deep learning models during clinical tasks.

Adaptive Segmentation Mask Attack

Adversarial examples are malicious data points that force machine learning models to make mistakes during testing time [2].

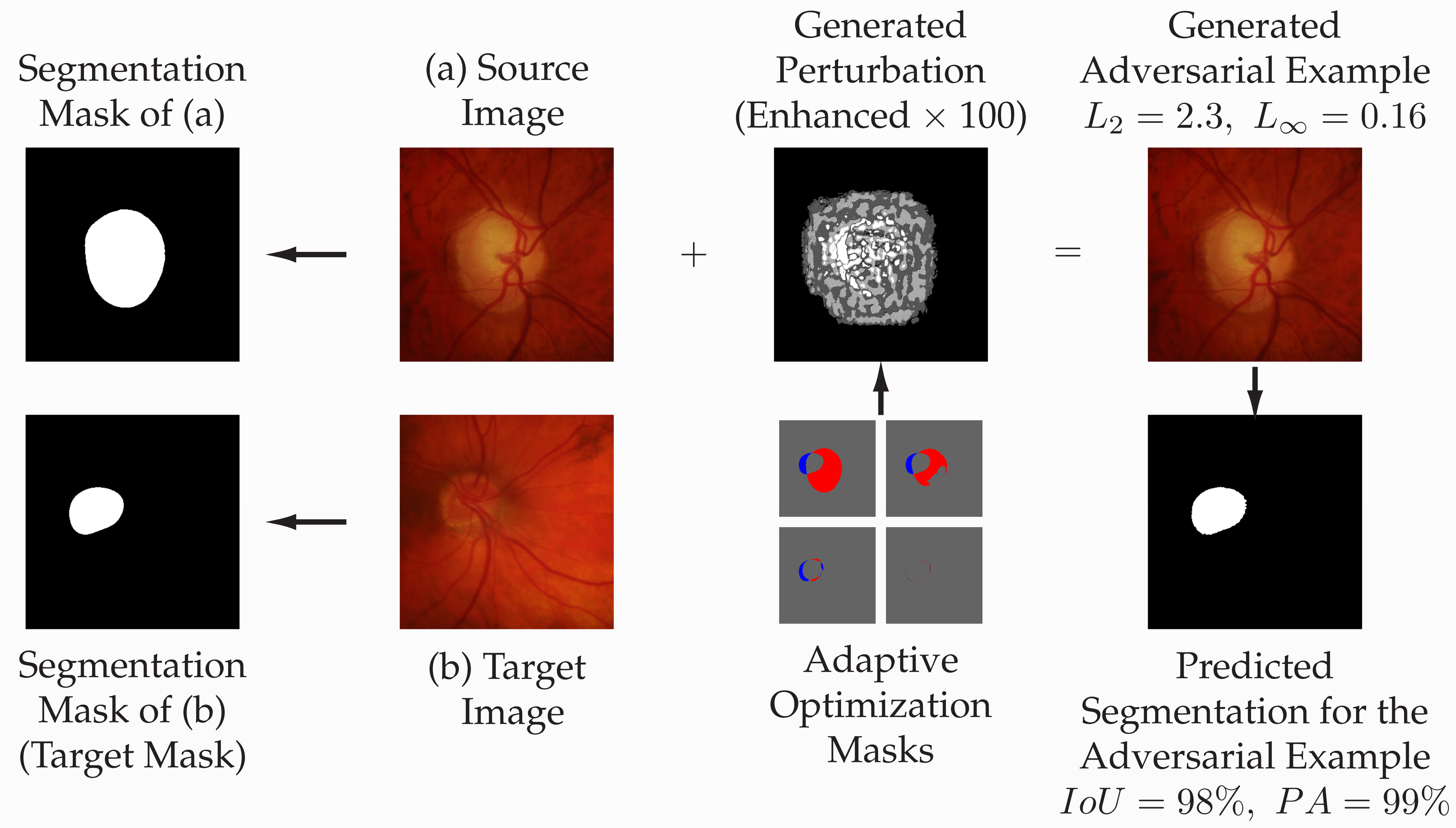


By introducing a novel algorithm for producing targeted adversarial examples for image segmentation problems, we expose the vulnerability of deep learning models for biomedical image segmentation to malicious data points. Our algorithm, named Adaptive Segmentation Mask Attack (ASMA), incorporates two techniques, namely, the use of (1) adaptive segmentation masks and (2) dynamic perturbation multipliers. The proposed attack is defined as follows:

$\mathbf{X}$: Input image.
$g(\theta, \mathbf{X})$: Forward pass from a neural network g with parameters θ using input $\mathbf{X}$.
$\mathbf{Y}^A$: Target (adversarial) mask.
$\mathbf{P}_n$: Added perturbation at nth iteration.

$$\text{minimize } \|\mathbf{X} - (\mathbf{X}+\mathbf{P})\|_2 \,,$$
$$\text{such that } \arg\max\big(g(\theta,(\mathbf{X}+\mathbf{P}))\big) = \mathbf{Y}^A \,, \quad (\mathbf{X}+\mathbf{P}) \in [0,1]^z \,,$$
$$\mathbf{P}_n = \sum_{c=0}^{M-1} \nabla_x \big(g(\theta,\mathbf{X}_n)_c \odot \mathbb{1}_{\{\mathbf{Y}^A = c\}} \odot \mathbb{1}_{\{\arg\max_M(g(\theta,\mathbf{X}_n)) \neq c\}}\big) \,.$$

ASMA is able to craft adversarial examples with 97% and 89% Intersection-over-Union (IoU) accuracy for the Glaucoma Dataset [3] and the ISIC Skin Lesion Dataset [4], respectively, with IoU measured between the predicted segmentation for a given adversarial example and the corresponding target mask. While doing so, our algorithm modifies the image so subtly that the perturbations, for the most part, are not visible to the bare eye.



Using ASMA, results obtained for the two above-mentioned biomedical datasets (mean and standard deviation) are provided in the table below (PA denotes Pixel Accuracy).

	Glaucoma Dataset				ISIC Skin Lesion Dataset			
	Modification		Accuracy		Modification		Accuracy	
Optimization	L_2	L_∞	IoU	PA	L_2	L_∞	IoU	PA
ASMA	**2.47** ±1.05	0.17 ±0.09	**97**% ±2%	99% ±1%	**3.88** ±1.99	0.16 ±0.09	**89**% ±10%	98% ±1%

* The experiments presented above are conducted in white-box settings, using the U-Net architecture [5].

References

[1] Finlayson S.G., Chung H.W., Kohane, I.S., Beam A.L., *Adversarial Attacks Against Medical Deep Learning Systems*
[2] Szegedy C., Zaremba W., Sutskever I., Bruna J., Erhan D., Goodfellow I., Fergus R., *Intriguing Properties of Neural Networks*
[3] Pena-Betancor C., Gonzalez-Hernandez M., Fumero-Batista F., Sigut J., Medina-Mesa E., Alayon S., de la Rosa M., *Estimation of the Relative Amount of Hemoglobin in the Cup and Neuroretinal Rim using Stereoscopic Color Fundus Images*
[4] Gutman D., Codella N., Celebi M., Helba B., Marchetti M., Mishra N., Halpern A., *Skin Lesion Analysis toward Melanoma Detection*
[5] Ronneberger O., Fischer P., Brox T., *U-Net: Convolutional Networks for Biomedical Image Segmentation*

GHENT UNIVERSITY

Center for Biotech Data Science