



## PERTURBATION ANALYSIS OF ADVERSARIAL **ATTACKS IN THE SPATIAL DOMAIN** Utku Ozbulak<sup>1,2</sup>, Arnout Van Messem<sup>1,3</sup>, Wesley De Neve<sup>1,2</sup>

<sup>1</sup> Center for Biotech Data Science, Ghent University Global Campus, Incheon, South Korea <sup>2</sup>Department for Electronics and Information Systems, Ghent University, Ghent, Belgium <sup>3</sup>Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium

## What are Adversarial Examples?













Average model transferability and perturbation rate of 1000 low-confidence adversarial examples generated with multiple methods.

**Transferability from ResNet50 to Other Models** 

- Our analysis show that each technique follows a unique perturbation pattern even though they
- Rate of transferability for adversarial examples
- The rate of transferability is higher when the perturbations focus on the main object of the

- Finding the best attack that produces robust adversarial with minimal perturbation is not easy due to non-convex nature of the optimization.
- Detecting adversarial examples still stands as one of the biggest challenges in the field with no effective solution yet.

<u>Technique</u>	<u>Perturbation</u>	<u>AlexNet</u>	<u>VGG16</u>	<u>ResNet152</u>
<b>GA</b> [2]	6.2%	36%	21%	13%
L-BFGS[3]	5.7%	32%	20%	9%
I-FGS <sub>[4]</sub>	6.1%	35%	20%	12%
C&W[5]	5.7%	51%	38%	25%

Even though multiple defense mechanisms are proposed, when the defense is incorporated into the optimization in a white-box attack, it is trivial to generate adversarial examples that bypass the proposed defense.

[1] K. He, X. Zhang, S. Ren, J. Sun. *ImageNet Classification with Deep Convolutional Neural Networks* 

[2] D. Erhan, Y. Bengio, A. Courville, P. Vincent. *Visualizing Higher-layer Features of a Deep Network* 

[2] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, R. Fergus. Intriguing Properties of Neural Networks

[4] A. Kurakin, I. J. Goodfellow, S. Bengio. Adversarial Examples in the Physical World

[5] N. Carlini, D. Wagner. *Towards Evaluating the Robustness of Neural Networks* 

utku.ozbulak@ugent.be github.com/utkuozbulak Ghent University Global Campus, Korea