

PERTURBATION ANALYSIS OF ADVERSARIAL ATTACKS IN THE SPATIAL DOMAIN

Utku Ozbek^{1,2}, Arnout Van Messem^{1,3}, Wesley De Neve^{1,2}

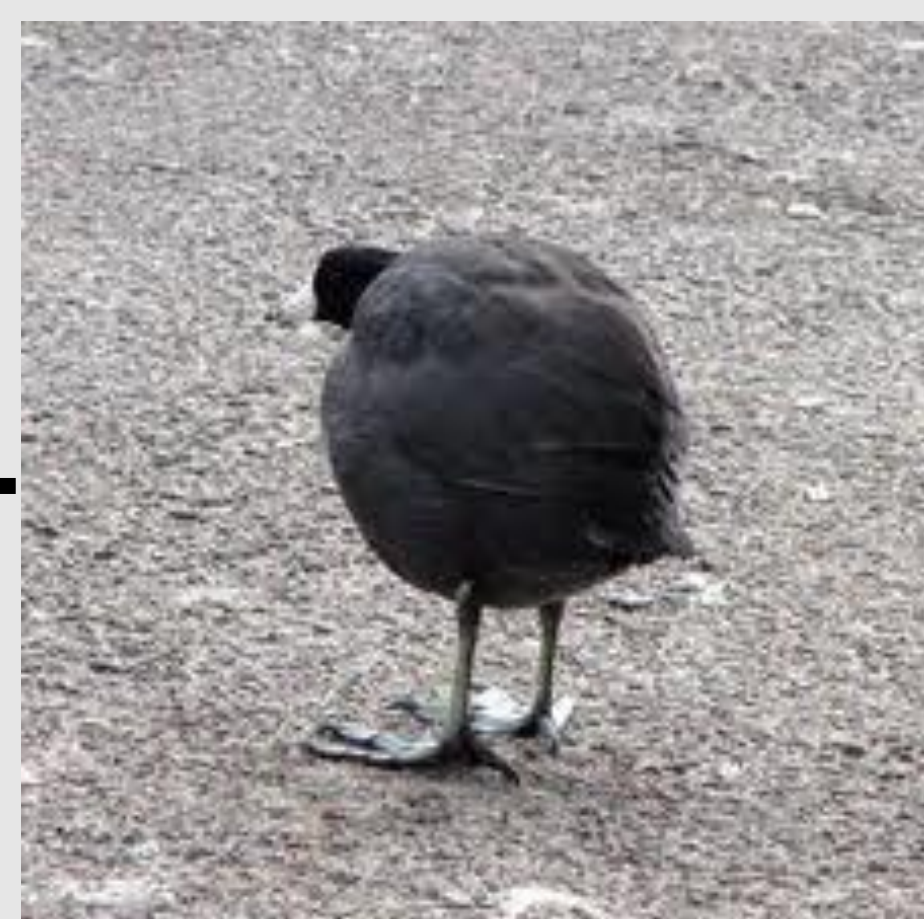
¹ Center for Biotech Data Science, Ghent University Global Campus, Incheon, South Korea

² Department for Electronics and Information Systems, Ghent University, Ghent, Belgium

³ Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium

What are Adversarial Examples?

Below experiments conducted on ResNet50 [1]



Original Image
Predicted as: Bird
Confidence: 96%



Fooling Image
Predicted as: Bird
Confidence: 100%

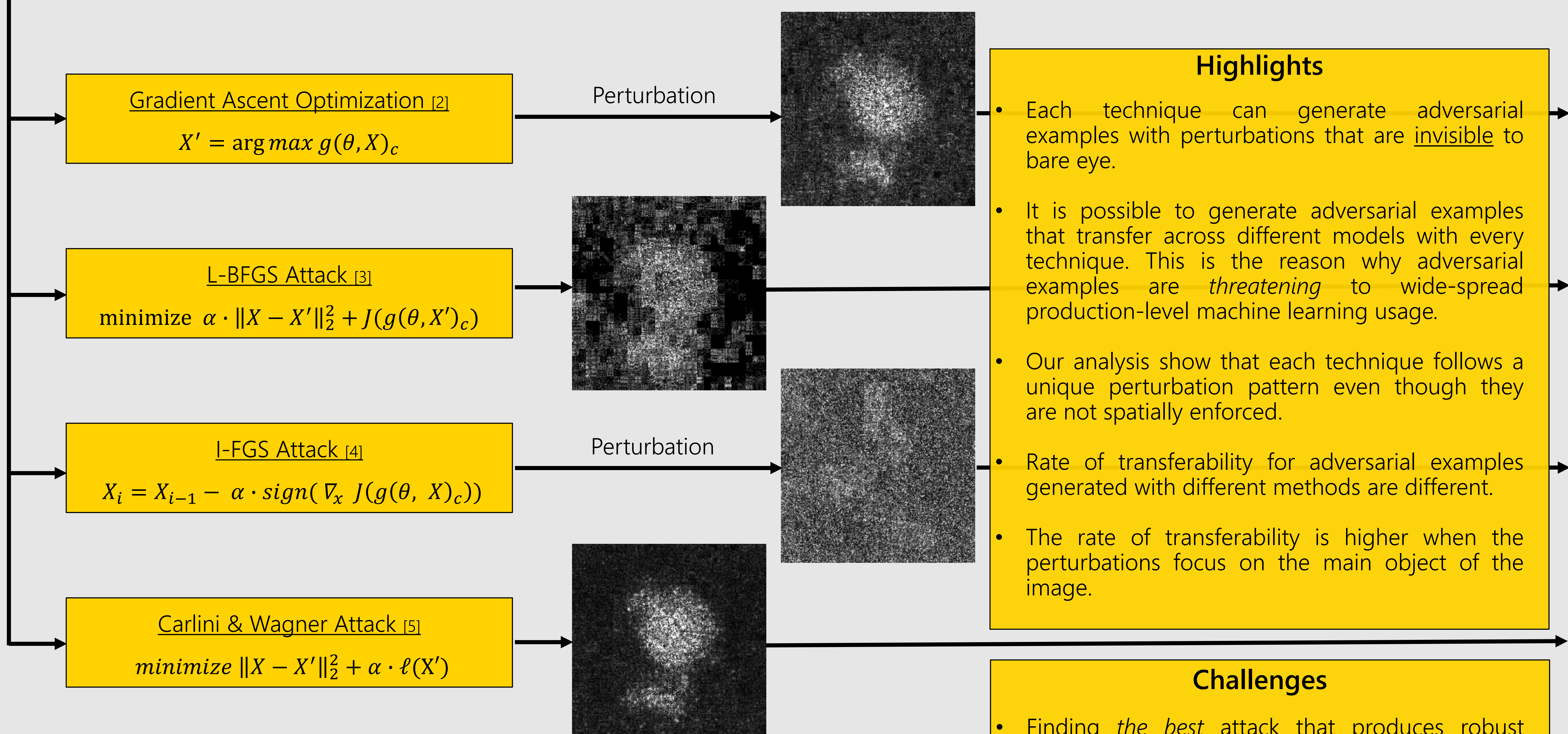


Adversarial Image (Maximization)
Predicted as: Rat
Confidence: 99%



Adversarial Image (Minimization)
Confidence of Bird: 7.6e-14%
(Least probable outcome)

Perturbation Patterns of Adversarial Example Generation Techniques



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

Technique	Perturbation	Transferability from ResNet50 to Other Models		
		AlexNet	VGG16	ResNet152
GA _[2]	6.2%	36%	21%	13%
L-BFGS _[3]	5.7%	32%	20%	9%
I-FGS _[4]	6.1%	35%	20%	12%
C&W _[5]	5.7%	51%	38%	25%

[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*

[3] 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*