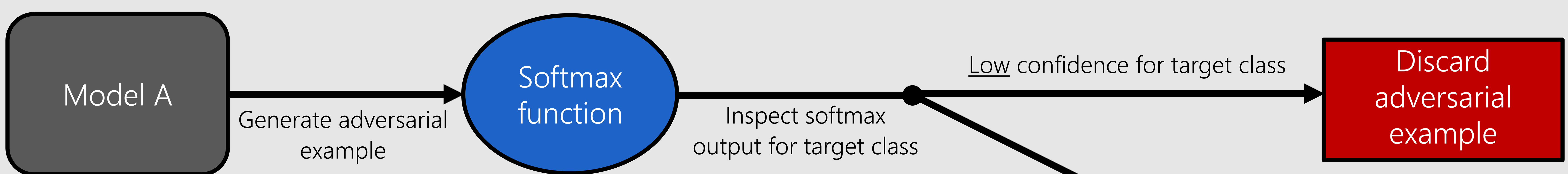
HOW THE SOFTMAX OUTPUT IS MISLEADING FOR EVALUATING THE STRENGTH OF ADVERSARIAL EXAMPLES

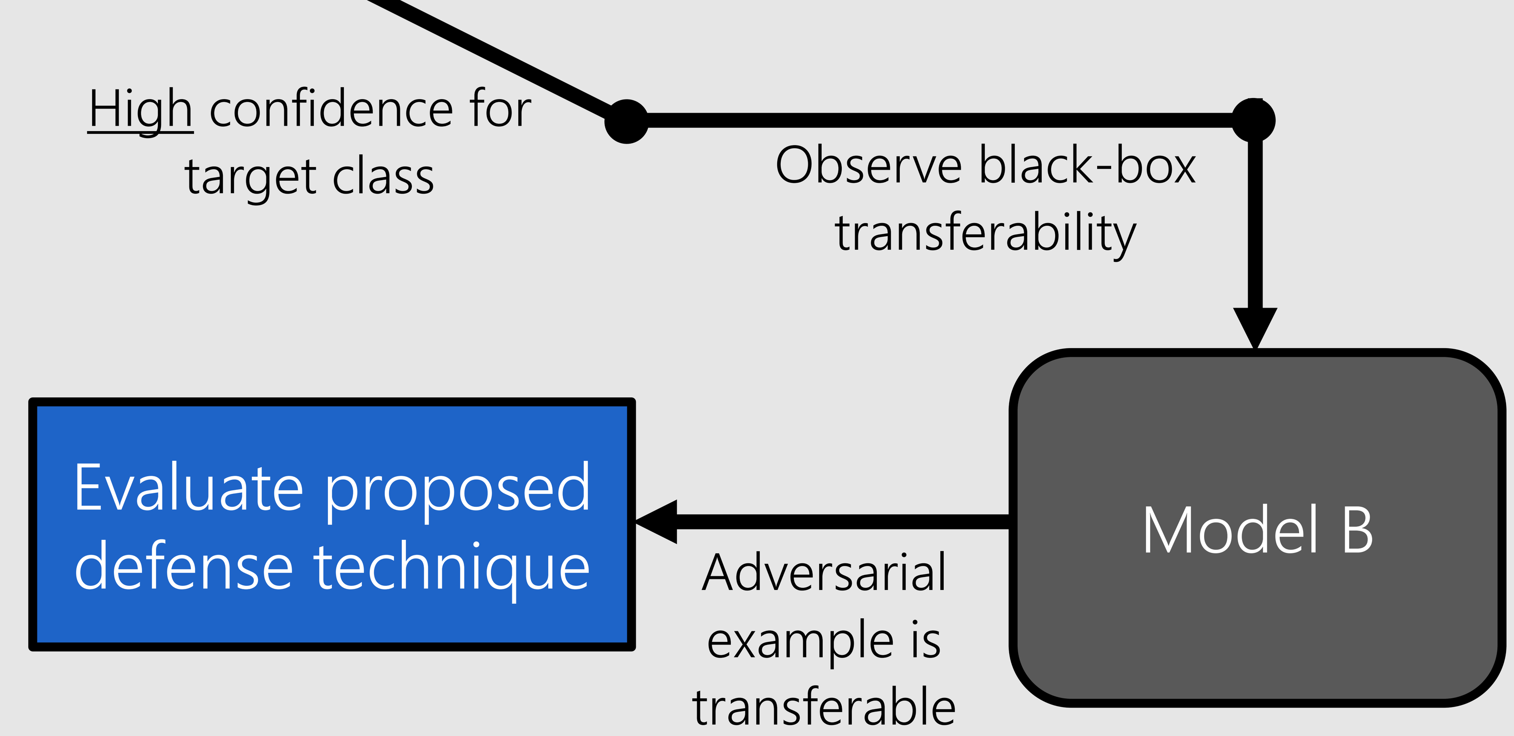
Utku Ozbulak[1,2] Wesley De Neve[1,2] Arnout Van Messem[1,3]

[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

A Common Method to Select *Strong* Adversarial Examples in Black-box Settings

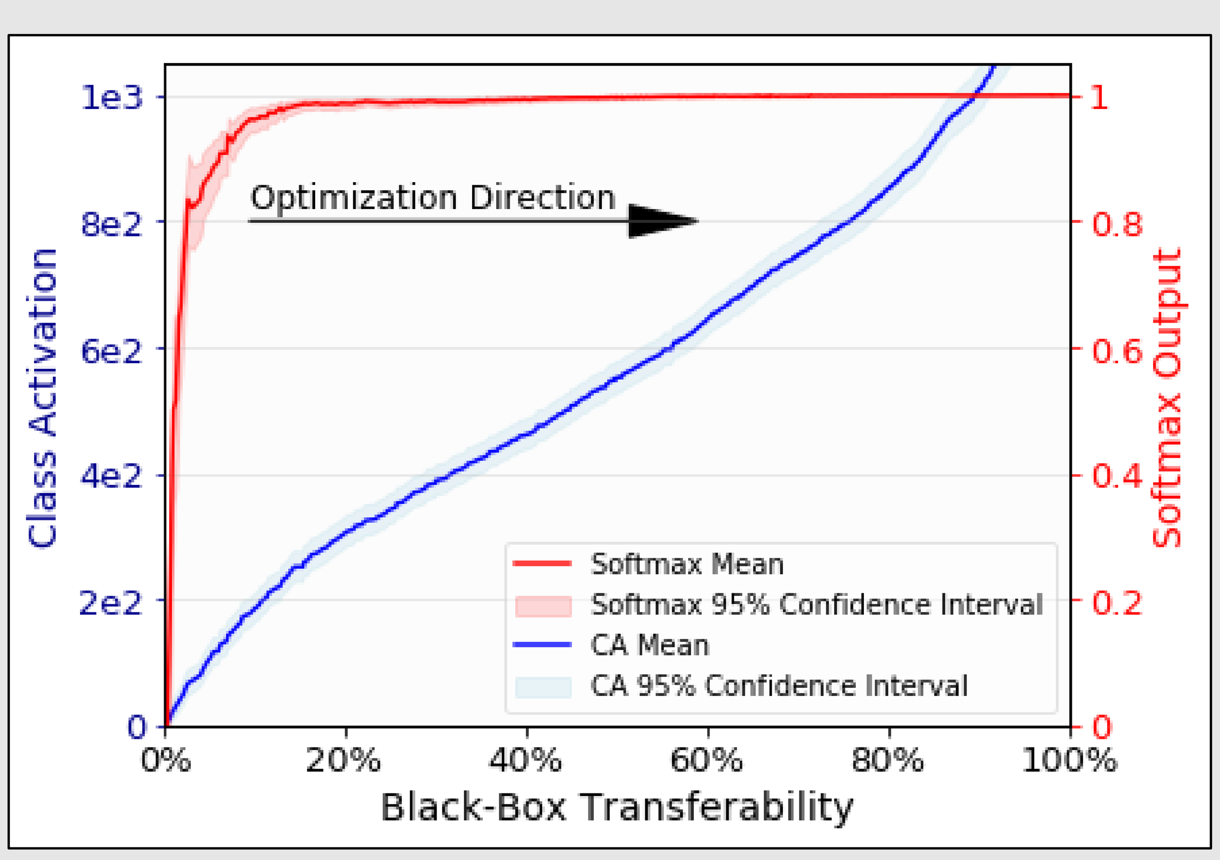


- Given a model for generating adversarial examples, a common way to select a *strong* adversarial example in black-box settings is to look for high confidence in the softmax output for the target class.
- However, based on black-box transferability, the softmax output does not correlate with the strength of an adversarial example.
- The softmax output for adversarial examples
  - is misleading when the optimization targets a single class (over-optimized adversarial examples).
  - can be easily manipulated with multi-target attacks (multi-class optimized adversarial examples).



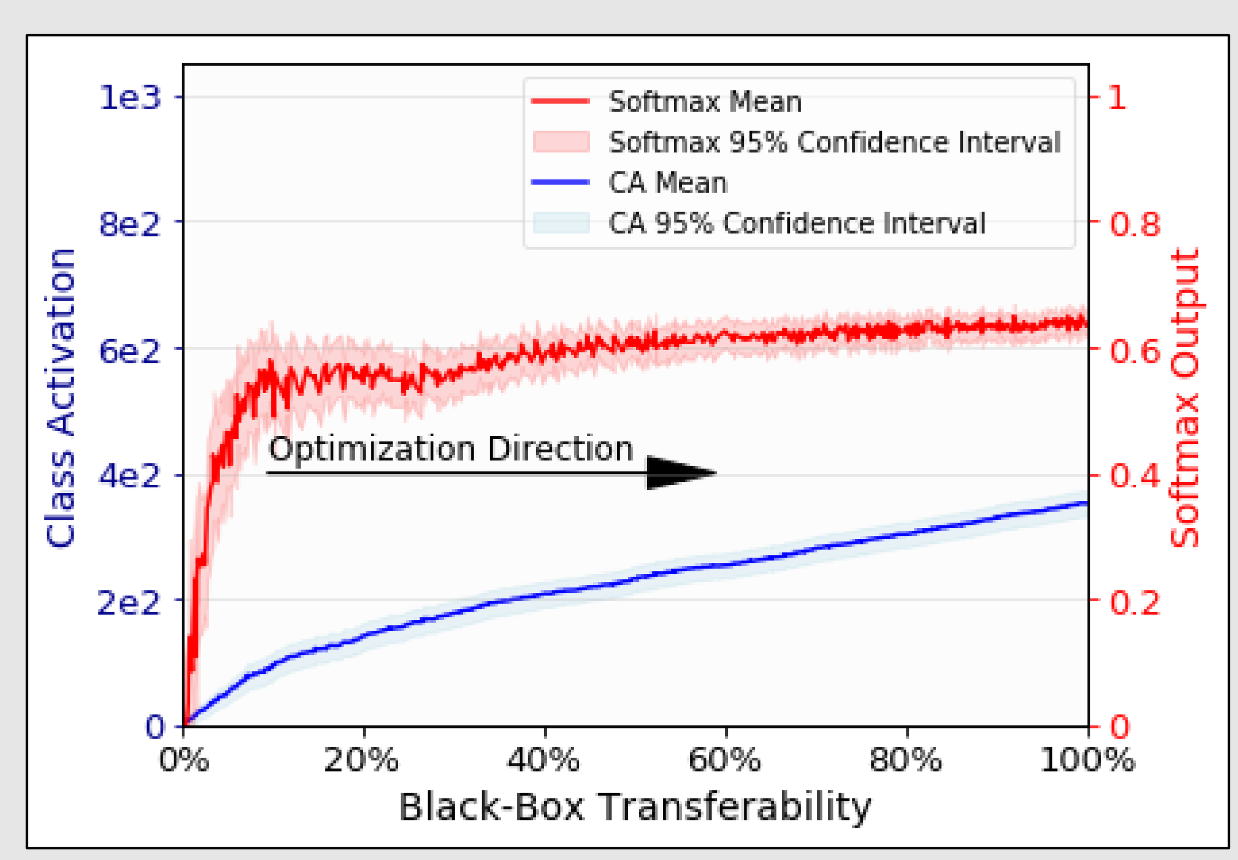
Over-optimized Adversarial Examples

- Generated by iteratively maximizing a single target class (similar to IFGS [1]): $X_{i+1} = X_i - \alpha \nabla_x g(\theta, X_i)_c$.
- Forces the softmax output to shoot off to 100% immediately, making it impossible to detect whether the adversarial example is strong or not based on the softmax output.



Multi-class Optimized Adversarial Examples

- Generated by iteratively maximizing multiple classes (similar to CW [2]): $X_{i+1} = X_i - \alpha \nabla_x g(\theta, X_i)_c - \beta \nabla_x g(\theta, X_i)_d$.
- Forces the softmax output to stay idle with a value less than 100%, creating misleading results when the output of the softmax is observed.



The graphs above show the mean target class activations (logit values) and their corresponding softmax output as a function of black-box transferability from VGG-16 [3] to ResNet-50 [4], for a total of 2000 adversarial examples.

[1] A. Kurakin, I. J. Goodfellow, S. Bengio. *Adversarial Examples in the Physical World*
[2] N. Carlini, D. Wagner. *Towards Evaluating the Robustness of Neural Networks*
[3] K. Simonyan, A. Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*
[4] K. He, X. Zhang, S. Ren, J. Sun. *ImageNet Classification with Deep Convolutional Neural Networks*

utku.ozbulak@ugent.be
(not present due to visa issues)
github.com/utku.ozbulak

Ghent University Global Campus, Korea