UNIVERSITY of RWANDA

ACEIoT

ACEIOT Research Seminar

Date: 08th May 2023

Venue: UR_CST /Einstein building/EAIFR Conference room
Time: 10h00 AM Kigali Time

Mode: Physical

Research on adversarial attack to deep neural networks

Prof. Jie Yang
Shanghai Jiao Tong University

China

Bio:

Jie Yang received a bachelor's degree in Automatic Control in Shanghai Jiao Tong University, where a master's degree in Pattern Recognition & Intelligent System was achieved three years later. In 1994, he received Ph.D. at Department of Computer Science, University of Hamburg, Germany. Now he is the Professor and Director of Institute of Image Processing and Pattern recognition in Shanghai Jiao Tong University. He is the principal investigator of more than 30 national and ministry scientific research projects in image processing, pattern recognition, data mining, and artificial intelligence. He has published six books, more than five hundreds of articles in national or international academic journals and conferences. Google citation over 17000, H-index 68. Up to now, he has supervised 5 postdoctoral, 36 doctors and 66 masters, awarded eight research achievement prizes from ministry of Education, China and Shanghai municipality. Two Ph.D. dissertation he supervised was evaluated as "National Best Ph.D. Dissertation" in 2009, in 2017, in 2019. He has owned 48 patents.

Abstract:

Despite the great success of deep neural networks, the adversarial attack can cheat some well-trained classifiers by small permutations. In this talk, we address two points.

•We propose a specific type of adversarial attack that can cheat classifiers by significant changes. Statistically, the existing adversarial attack increases Type II error and the proposed one aims at Type I error, which are hence named as Type II and Type I adversarial attack, respectively. To implement the proposed attack, a supervised variation autoencoder is designed and then the classifier is attacked by updating the latent variables using gradient information. Besides, with pretrained generative models, Type I attack on latent spaces is investigated as well. The research was published in TPAMI in 2021.

•The existing adversarial attacks have high success rates only when the information of the victim DNN is well-known or could be estimated by the structure similarity or massive queries. We propose Attack on Attention (AoA), a semantic property commonly shared by DNNs. AoA enjoys a significant increase in transferability when the traditional cross entropy loss is replaced with the attention loss. Since AoA alters the loss function only, it could be easily combined with other transferability-enhancement techniques and then achieve SOTA performance. We apply AoA to generate 50000 adversarial samples from ImageNet validation set to defeat many neural networks, and thus name the dataset as DAmageNet. The research was published in TPAMI in 2022.