MACQUARIE'S DISTINGUISHED DSAI SEMINARS

MACQUARIE University

PROF JIE YANG

SHANGHAI JIAOTONG UNIVERSITY

Research on adversarial attack and robustness of deep neural networks

Abstract:

Despite the great success of deep neural networks, the adversarial attack can cheat some well-trained classifiers by small permutations. 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 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. Some recent research results on the robustness of deep neural networks will also be presented. Three papers about above topics have been published in TPAMI and PR in recent years.

Biography:

Prof Jie Yang received a bachelor's degree in Automatic Control in Shanghai Jiao Tong University (SJTU), 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 19000. H-index 65. Up to now, he has supervised 5 postdoctoral, 46 doctors and 70 masters, awarded six research achievement prizes from ministry of Education, China and Shanghai municipality. He has owned 48 patents. Three Ph.D. dissertation he supervised was evaluated as "National Best Ph.D. Dissertation" in 2009, in 2017, in 2019. He has been chairman and keynote speaker of more than 10 international conferences.

Macquarie's School of Computing and DataX Consilience Research Centre jointly launch the Macquarie's Distinguished AI Seminar series for distinguished speeches on AI frontiers, recent advances, and best practices by renowned thought-leaders, researchers and practitioners. The Distinguished AI Seminar series aims to promote cross-disciplinary AI research, innovation, engagement and applications, and bridge gaps between cutting-edge research and impactful applications in AI, data science, machine learning and other relevant areas.

DATE: Friday, Oct 6, 2023

TIME: 1.30 pm - 2.30pm

LOCATION: 4 Research

Park Drive, Level 2, 288,

Macquarie University

ZOOM ID: 89587728714

AIsummits.org
mqDataX.net
comp.mq.edu.au
mq.edu.au

AI.mq@mq.edu.au

MACQUARIE University SYDNEY·AUSTRALIA

DATAX RESEARCH CENTRE

DSAI@MQ