DSAI
DATA SCIENCE & ARTIFICIAL INTELLIGENCE
dsai.etf.unsa.ba

RESEARCHES ON THE DEFENSES AND OUT-OF-DISTRIBUTION DETECTION IN TRUSTWORTHY DEEP LEARNING

RESEARCH SEMINAR

JULY 20 | 2026 11:00 AM - 12:00 PM

Prof. Dr. Jie Yang

Professor and Director of Institute of Image Processing and Pattern recognition in Shanghai Jiao Tong University

Prof. Dr. 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.

About seminar: The rapid advancement of deep learning has had a transformative effect on the development of technology and society across a multitude of sectors. In safety-critical contexts, the potential for neural network models to produce unreliable outputs in response to "malicious" or "unanticipated" inputs poses a severe risk. This talk delves into the output reliability from neural network models within the domain of trustworthy deep learning. 1) Inputs that involve pixel perturbations, exemplified by adversarial examples, w.r.t the task of adversarial and certified robustness; 2) Inputs that represent distribution shifts, exemplified by Out-of-Distribution (OoD) data, w.r.t the task of out-of-distribution detection. We introduce a novel strategy of model augmentation, adopt a multi-head neural network structure, and pose diversity constraints related to adversarial robustness into the model parameters. We adopt a multi- head neural network structure, use the ensemble of multiple heads in place of the ensemble of multiple neural networks, which significantly reduces the computational load in both training and certification phases. We propose that the non-linearity in InD and OoD data hinders PCA from learning a subspace that fully embodies their diversities. We propose a mode ensemble method that not only enhances detection performance but also significantly reduces the performance variance among independent modes.We propose performing linear dimension reduction on the gradient using a designated subspace that comprises principal components. Papers about above topics have been published in TPAMI, IJCV, PR, NIPS in recent years.

Location: **Bit Alliance Lab (3-46), III floor**
Faculty of Electrical Engineering,
University of Sarajevo

REGISTER HERE

Also available online via zoom!