Substitute Training for Adversarial Attacks - Is Real Training Data Really Necessary?

Abstract: Recent study shows machine learning models are extremely vulnerable to adversarial attacks. Substitute attacks, typically black-box ones, employ pre-trained models to generate adversarial examples. It is generally accepted that substitute attacks need to acquire a large amount of real training data combined with model-stealing methods to obtain a substitute model. However, the real training data may be difficult (if not impossible) to be obtained for some practical tasks, e.g., in medical or financial sectors. As the first trial study, the talk will present our recently developed data-free model-stealing method for substitute training that does not require any real training data. The experimental results demonstrate that the substitute models produced by the proposed method without any real training data can achieve competitive performance against the baseline models trained by the same training set as in attacked models.

Speaker:

Ce Zhu is currently a Professor with the School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu, China. His research interests include image/video coding and communications, 3D video, visual analysis and understanding, visual perception and applications. He has served on the editorial boards of a few journals, including as an Associate Editor of *IEEE Transactions on Image Processing*, *IEEE Transactions on Circuits and Systems for Video Technology*, *IEEE Transactions on Broadcasting*, *IEEE Signal Processing Letters*, and *IEEE Communications Surveys and Tutorials*. He has also served as a Guest Editor of a few special issues in international journals, including as a Guest Editor in the *IEEE Journal of Selected Topics in Signal Processing*. He is a Fellow of the IEEE, and an IEEE CASS Distinguished Lecturer (2019-2020).