香港城市大學
City University of Hong Kong
混沌及複雜網絡研究中心
Centre for Chaos and Complex Networks

THE HONG KONG
POLYTECHNIC UNIVERSITY
Department of Electronic and Information Engrg.
香港理工大學 電子及資訊工程學系

IEEE Technical Co-sponsor: IEEE Hong Kong Section
Robotics and Automation/Control Systems Joint Chapter

Jointly presents

SEMINAR SERIES ON COMPLEX SYSTEMS, NETWORKS, CONTROL AND APPLICATIONS

Attacks and Defenses on Machine Learning Services

Mr. Huadi Zheng
The Hong Kong Polytechnic University, Hong Kong

Date and Time: Friday, 1 November 2019, 4:30pm – 5:30pm

Venue: Room **CD634**, Hong Kong Polytechnic University
Reception starts at 4:15pm
(Language: **English**)

Abstract

Service providers train machine learning models using large datasets owned or acquired by themselves, and use these models to offer online services, such as face and voice recognition, through a public prediction API. However, a prediction API call, which consists of a query and its response, can be vulnerable to adversarial attacks that disclose the internal states of these models. Particularly, a model extraction attack is able to restore important model parameters using the rich information (e.g., model type, prediction confidence) provided by the prediction API. Once the model is extracted, an adversary can further apply model inversion attack to learn the proprietary training data, compromising the privacy of data contributors. Another follow-up attack on the extracted model is evasion attack, which avoids a certain prediction result by modifying its query. As countermeasures, researchers have proposed to reduce the rich API output, such as hiding the precise confidence level of the prediction response. Adversarial training is also proposed to increase the robustness against adversarial samples. This talk gives a brief review of these attacks and defences on machine learning services. Particularly, a boundary differentially private layer will be presented to introduce active defence against model extraction attacks.

About the Speaker

Huadi Zheng receives the BEng degree in software engineering from the School of Data and Computer Science, Sun Yat-sen University, Guangdong, in 2012. Currently he is pursuing a PhD degree in the Department of Electronic and Information Engineering, Hong Kong Polytechnic University. His research interests include mobile side-channel security, data privacy and machine learning security.

Inquiry: G. Ron Chen (gchen@ee.cityu.edu.hk) or Michael C. K. Tse (encktse@polyu.edu.hk)