Temple University

Spring 2022 Colloquium Series
Department of Computer and Information Sciences

Fostering Trustworthiness in Machine Learning via Robust and Automated Model Interpretation

Mengdi Huai
University of Virginia

Monday, January 24th, 11am
Zoom Link: https://temple.zoom.us/j/96541275510

Abstract: Machine learning models have been widely applied in real world to build intelligent systems (e.g., self-driving cars, intelligent recommendation systems, and clinical decision support systems). However, traditional machine learning models mainly focus on optimizing accuracy and efficiency, and they fail to consider how to foster trustworthiness in their design. In practice, machine learning models are suffering a crisis of trust when they are applied in real-world applications due to the lack of transparency behind their behaviors. The concern about the "black box" nature of machine learning models makes decision makers reluctant to trust the predicted results, especially when these models are used for making critical decisions (e.g., medical disease diagnosis). In this talk, I will introduce my research efforts towards the goal of making machine learning trustworthy. Specifically, I will discuss how to foster trustworthiness in machine learning via robust and automated model interpretation. I will first describe my recent research on the security vulnerability of model interpretation methods for deep reinforcement learning (DRL) and introduce two malicious attack frameworks that can significantly alter the interpretation results while incurring minor damage to the performance of the original DRL model. Then, I will present an automated and robust model interpretation framework, which can not only automatically generate the concept-based explanations for the predicted results but also provide certified robustness guarantees for the generated explanations.

Bio: Mengdi Huai is a Ph.D. candidate in the Department of Computer Science at the University of Virginia. Her research interests lie in the areas of data mining and machine learning, with a current focus on developing novel techniques to build trustworthy learning systems that are explainable, robust, private, and fair. Mengdi is also interested in designing effective data mining and machine learning algorithms to deal with complex data with both strong empirical performance and theoretical guarantees. Her research work has been published in various top-tier venues, such as KDD, AAAI, IJCAI, NeurIPS, and TKDD. Mengdi received multiple prestigious awards from the University of Virginia for her excellence in research, including the Sture G. Olsson Fellowship in Engineering and the John A. Stankovic Research Award. Her recent work on malicious attacks against model interpretation won the Best Paper Runner-up of KDD2020. Mengdi was selected as one of the Rising Stars in EECS at MIT. She was also selected as one of the Rising Stars in Data Science at UChicago.