CS Seminar

Robust and Sparse Interpretation of Deep Neural Networks

Dr. Farzan Farnia
Department of Computer Science and Engineering
The Chinese University of Hong Kong

Date:
Feb 24, 2023
Friday
4:00pm

Venue:
Room 328
Chow Yei Ching Bldg
The University of Hong Kong

Join zoom meeting
https://hku.zoom.us/j/97844793304?pwd=dlpwaUF0TzQ1OS9HUFZFRWl2SmpaZz09
Meeting ID: 978 4479 3304
Password: 782973

Abstract:

Explaining the predictions of deep neural networks has been a topic of great interest in the machine learning community. In this talk, we will focus on gradient-based saliency maps for interpreting neural network models. We will discuss the vulnerability of standard gradient-based interpretation schemes to input perturbations, and introduce MoreauGrad as an interpretation scheme based on a neural network's Moreau envelope. We prove the certifiable robustness of MoreauGrad to norm-bounded input perturbations, and subsequently propose a sparse version of MoreauGrad by applying L1-norm regularization to its formulation. We discuss the robustness properties of Sparse MoreauGrad and display the visual performance of our proposed interpretation scheme in application to standard image datasets.

About the Speaker:

Farzan Farnia is an Assistant Professor of Computer Science and Engineering at The Chinese University of Hong Kong. Prior to joining CUHK, he was a postdoctoral research associate at the Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, from 2019-2021. He received his master's and PhD degrees in electrical engineering from Stanford University and his bachelor's degrees in electrical engineering and mathematics from Sharif University of Technology. At Stanford, he was a graduate research assistant at the Information Systems Laboratory advised by David Tse. Farzan's research interests span statistical learning theory, information theory, and convex optimization.

All are welcome!
For enquiries, please call 2859 2180 or email enquiry@cs.hku.hk
Department of Computer Science
The University of Hong Kong