Title: Explaining Neural Networks: A Causal Perspective

Speaker: Dr. Vineeth Balasubramanian, IIT Hyderabad

Time, Place: 15.00, 4/7/2019, KD-102

Abstract: As neural network (deep learning) models get absorbed into real-world applications each day, there is an impending need to explain the decisions of these neural network models. This talk will begin with an introduction to the need for explaining neural network models, summarize existing efforts in this regard, as well as present a few of our efforts in this direction. In particular, while existing methods for neural network attributions (for explanations) are largely statistical, we propose a new attribution method for neural networks developed using first principles of causality (to the best of our knowledge, the first such). The neural network architecture is viewed as a Structural Causal Model, and a methodology to compute the causal effect of each feature on the output is presented. With reasonable assumptions on the causal structure of the input data, we propose algorithms to efficiently compute the causal effects, as well as scale the approach to data with large dimensionality. We also show how this method can be used for recurrent neural networks. We report experimental results on both simulated and real datasets showcasing the promise and usefulness of the proposed algorithm. This work was presented this year at ICML 2019.

About speaker:
Vineeth N Balasubramanian is an Associate Professor in the Department of Computer Science and Engineering at the Indian Institute of Technology, Hyderabad. His research interests include deep learning, machine learning, and computer vision. His research has been published at premier peer-reviewed venues including ICML, CVPR, ICCV, KDD, ICDM, IEEE TPAMI and ACM MM. His PhD dissertation at Arizona State University on the Conformal Predictions framework was nominated for the Outstanding PhD Dissertation at the Department of Computer Science. He is an active reviewer/contributor at many conferences such as NeurIPS, CVPR, ICCV, AAAI, IJCAI, ACM MM, as well as journals including IEEE TPAMI, IEEE TNNLS, JMLR and Machine Learning. He currently serves as the Secretary of the AAAI India Chapter. For more details, please see https://iith.ac.in/~vineethnb/.