Computer Science & Engineering Spring Lecture Series 2023

UCONN
UNIVERSITY OF CONNECTICUT

Speaker: Hanjie Chen, PhD Candidate, University of Virginia

Date: Monday, March 20, 2023
Time: 12:00-1:00 EST
Location: UConn Library Conference Room 1102
Webex: https://uconn-cmr.webex.com/meet/cdc19010

Title: Bridging the Trustworthy Gap between AI and Humans: Interpretation Techniques for Modern NLP

Abstract: Neural network models have been pushing computers' capacity limit on natural language understanding and generation while lacking interpretability. The black-box nature of deep neural networks hinders humans from understanding their predictions and trusting them in real-world applications. In this talk, I will introduce my effort in bridging the trustworthy gap between models and humans by developing interpretation techniques, which cover three main phases of a model life cycle—training, testing, and debugging. I will demonstrate the critical values of integrating interpretability into every state of model development: (1) making model prediction behavior transparent and interpretable during training; (2) explaining and understanding model decision-making on each test example; (3) diagnosing and debugging models (e.g., robustness) based on interpretations. I will discuss future directions on incorporating interpretation techniques with system development and human interaction for long-term trustworthy AI.

Bio: Hanjie Chen is a Ph.D. candidate in Computer Science at the University of Virginia. Her research interests lie in Trustworthy AI, Natural Language Processing (NLP), and Interpretable Machine Learning. She is a recipient of the Carlos and Esther Farrar Fellowship and the Best Poster Award at the ACM CAPWIC 2021. Her work has been published at top-tier NLP/AI conferences (e.g., ACL, AAAI, EMNLP, NAACL) and selected by the National Center for Women & Information Technology (NCWIT) Collegiate Award Finalist 2021. Besides, as the primary instructor, she co-designed and taught a cross-listed course, CS 4501/6501 Interpretable Machine Learning, at UVA. Her effort in teaching was recognized by the UVA CS Outstanding Graduate Teaching Award and University-wide Graduate Teaching Awards Nominee (top 5% of graduate instructors).