University of Pittsburgh | School of Computing and Information

Dissertation Defense

Doctor of Philosophy in Computer Science

"**Multimodal Knowledge Integration for Object Detection and Visual Reasoning**"
by **Keren Ye**

Date: July 8, 2021
Time: 1:00pm – 3:00pm
Place: https://pitt.co1.qualtrics.com/jfe/form/SV_cGU6h6CDCHs0dLg

Committee:

- Adriana Kovashka, Assistant Professor, Department of Computer Science, School of Computing and Information, University of Pittsburgh
- Diane Litman, Professor, Department of Computer Science, School of Computing and Information, University of Pittsburgh
- Milos Hauskrecht, Professor, Department of Computer Science, School of Computing and Information, University of Pittsburgh
- Daqing He, Professor, Department of Informatics and Networked Systems, School of Computing and Information, University of Pittsburgh
- Seong Jae Hwang, Assistant Professor, Department of Computer Science, School of Computing and Information, University of Pittsburgh

Abstract:
We humans still perceive and reason in a different way than artificial intelligence models. We witness, we listen, we touch, we understand the world via multi-modal sensing, while machine models rely only on a single or a few modalities and ignore abundant information in nature. In this thesis, we explore techniques for reducing the perception gap between machines and humans. First, we incorporate information from text, audio, motion, external knowledge bases, for training computer vision models. We find that data inputs from more extensive channels provide complementary information to improve models. Second, we study how multimodal inputs can be fully utilized. We argue that most existing deep learning methods are prone to use only a proportion of input features, which causes the resulting models to be biased. We propose robust training to overcome the issue. Third, we extend the benefits of multi-modal information to the supervision signals instead of the inputs, by learning a weakly supervised detection model from the natural supervision of textual captions. We propose two directions to extend this research. On the one hand, we argue that learning a detection model does not fully unleash the power of using text modality as supervision. With the help of NLP constituency parsing, it is possible to determine also the properties and relations of visual objects. On the other hand, we plan to explore the use of videos, by extracting supervision from audio and motion features.