Frontier Topics in Vision and Language: Spring 2021 Seminar Series
https://asu-active-perception-group.github.io/seminar

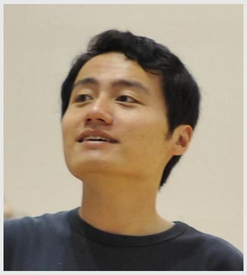
Learning Visual Knowledge from Image-Text Pairs

Monday, March 29, 2021, 16:30 MST
Via Zoom: https://asu.zoom.us/j/89817732811

Abstract

Images and their text descriptions (i.e., captions) are readily available in great abundance over the Internet, creating a unique opportunity to develop AI models for image and text understanding. Consequently, learning from these image-text data has received a surging interest from the vision and AI community. An image contains millions of pixels capturing the intensity and color of a visual scene. Yet the same scene can be oftentimes summarized using dozens of words in a natural language. How can we bridge the gap between visual and text data? And what can we learn from these image-text pairs? In this talk, I will describe our attempts to address these research questions, with a focus on learning visual knowledge from images and their captions.

First, I will talk about our early work on learning joint representations to match images and sentences and to further align regions with an image and phrases from the image caption. Our latest development demonstrates the learning of these representations with merely image-text pairs and without knowing region-phrase correspondences. Moving forward, I will present our recent work on learning to detect visual concepts (e.g., object categories) and their relationships (e.g., predicates) -- in the form of localized scene graphs, again from only image-sentence pairs. Lastly, I will describe our method that leverages image scene graphs to generate accurate, diverse, and controllable image captions. If time permits, I will briefly cover our efforts of wearable visual sensing and first person vision.



Yin Li
University of Wisconsin
Madison

Bio
Yin Li is an Assistant Professor in the Department of Biostatistics and Medical Informatics and affiliate faculty in the Department of Computer Sciences at the University of Wisconsin-Madison. Previously, he obtained his PhD in computer science from Georgia Tech and was a postdoctoral fellow in the Robotics Institute at the Carnegie Mellon University. His primary research focus is computer vision. He is also interested in the applications of vision and learning for mobile health. Specifically, his group develops methods and systems to automatically analyze human activities to address challenges related to healthcare. He has been serving as area chairs for the top vision and AI conferences, including CVPR, ICCV, ECCV and IJCAI. He was the co-recipient of the best student paper awards at MobiHealth 2014 and IEEE Face and Gesture 2015. His work was covered by MIT Tech Review, WIRED UK, New Scientist, BBC, and Forbes.

Host: Zhiyuan Fang, Yezhou Yang, Chitta Baral

The Active Perception Group explores robotic visual learning, tying together the fields of active vision, natural language processing and AI reasoning.

ASU Ira A. Fulton Schools of Engineering
Arizona State University

School of Computing, Informatics and Decision Systems Engineering