Department of Computer Science

Cross-Modal Learning for Video Understanding

SouYoung Jin
MIT

Hosted by Jing Li

Date: Friday, February 18, 2022
Seminar: 11:00 AM – 12:00 PM
Location: https://njit.webex.com/njit/j.php?MTID=ma49a195ff5270b2f018a9072e8ca814a

https://cs.njit.edu/seminars

Abstract:

Videos are good sources of knowledge about things we have not yet experienced. They also show many aspects of human life. Videos have multiple sources of sensory information. Building a video understanding system requires computer vision components, such as object detection and recognition, and knowledge from other domains such as spoken/natural language processing and cognitive science. Cross-modal learning is a way of learning that involves information obtained from more than one modality. In this talk, I will introduce two recent projects on cross-modal learning for video understanding. In particular, I will talk about the "Spoken Moments" project, where my collaborators and I collected spoken descriptions of 500K short videos, to capture natural and concise descriptions on a large scale. We designed the study to collect only descriptions of events that stood out in participants' memory, as we were particularly interested in the video content that human annotators pay attention to. Using pairs of video and spoken descriptions, we trained a model with a cross-modal learning architecture to understand the video content, leading to more human-like understanding. The model trained on the spoken moments generalizes very strongly to the other datasets. I will also present our approaches to model training and future projects in video understanding.

Bio:

Dr. SouYoung Jin is a postdoctoral associate at the Computer Science and Artificial Intelligence Laboratory (CSAIL) at the Massachusetts Institute of Technology (MIT), working with Dr. Aude Oliva in the Computational Perception & Cognition Lab. Her main research area is in computer vision, machine learning and cognitive science. She has also extensively collaborated with experts in related fields such as spoken/natural language processing. Her area of expertise is video understanding.

Dr. Jin earned her PhD in 2020 at the College of Information and Computer Sciences (CICS), University of Massachusetts, Amherst (UMass Amherst), where she researched how to improve face clustering in videos under Dr. Erik Learned-Miller in the Computer Vision Lab. She also completed an internship at Microsoft AI and Research with Dr. Lei Zhang in 2018. Dr. Jin was selected to participate in the prestigious Rising Stars 2019 -- an academic career workshop for women in EECS. She co-organized a workshop on Multi-Modal Video Analysis at ECCV 2020.