COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF MICHIGAN

DISSERTATION DEFENSE

Oana Ignat

Towards Human Action Understanding in Social Media Videos using Multimodal Models

Thursday, July 28, 2022
2:00 – 4:00pm
Virtual Event
Zoom – Passcode: 6vbTMdD9

ABSTRACT: Human action understanding is one of the most impactful and challenging tasks a computer system can do. Once a computer system learns how to interact with humans, it can assist us in our everyday life activities and significantly improve our quality of life.
Despite the attention it has received in fields such as Natural Language Processing and Computer Vision, and the significant strides towards accurate and robust action recognition and localization systems, human action understanding still remains an unsolved problem.
In this thesis, we introduce and analyze how models can learn from multimodal data, i.e, from what humans *say* and *do* while performing their everyday activities.
As a step towards endowing systems with a richer understanding of human actions in online videos, this thesis proposes new techniques that rely on the vision and language channels to address four important challenges: i) human action visibility identification in online videos, ii) temporal human action localization in online videos, iii) human action reason identification in online videos, and iv) human action co-occurrence identification.
We focus on the widely spread genre of lifestyle vlogs, which consist of videos of people performing actions while verbally describing them. We construct a dataset with crowdsourced manual annotations of visible actions, temporal action localization and action reason identification in online vlogs.
We propose a multimodal unsupervised model to automatically infer the reasons corresponding to an action presented in the video, a simple yet effective method to localize the narrated actions based on their expected duration, and a multimodal supervised classification model of action visibility in videos. We also perform ablations on how each modality contributes to solving the tasks and compare the multimodal models performance with the single-modalities models based on the visual content and vlog transcripts.
Finally, we present an extensive analysis of this data, which allows for a better understanding of how the language and visual modalities interact throughout the videos and pave the road for rich avenues for future work.

CHAIR: Prof. Rada Mihalcea