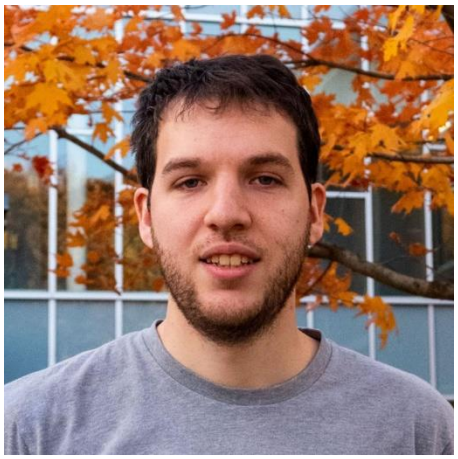
COMPUTER SCIENCE & ENGINEERING
UNIVERSITY OF MICHIGAN

DISSERTATION DEFENSE



Santiago Castro

Towards Video Understanding through Language in Real-life Settings

Monday, June 3, 2024
10:00am – 12:00pm
3725 Beyster
Hybrid – Zoom

ABSTRACT: Videos have become an integral part of our daily lives, with a rapidly growing number on YouTube, Netflix, and TikTok serving as testimony to their widespread popularity. Behind the simplicity of their interfaces and user experiences, the systems that power these products employ numerous video-understanding techniques, even for straightforward use cases such as finding a video on how to cook salmon. Despite the significant progress achieved in this area, there remains a gap between lab-setting capabilities and reality, as multiple phenomena are not adequately designed for realistic settings, causing various issues such as domain mismatches and the diverse way people interact in videos. My work aims to bridge this gap by enabling the understanding of video content in realistic settings.

The issues that make current video understanding research unsuitable for real life can be classified into data, methods, and evaluation. The data aspect is crucial since current research has predominantly overlooked real-life settings. I present new datasets and benchmarks for such domains: daily situations and in-the-wild scenarios. These benchmarks measure the effectiveness of new methods in these more realistic settings. Likewise, I introduce a novel framework that accounts for a typical yet understudied human behavior: sarcasm. Sarcasm is particularly suited to be studied in video since I show that leveraging what we see and hear allows one to understand it better. For the methods aspect, I consider a fundamental issue, which is the impracticality and lack of scalability of the traditional in-the-lab setting, tuning one model for each newly addressed task and domain. I propose a robust method that allows practitioners to employ a single model for novel tasks and domains with satisfactory performance. Additionally, I present a technique to improve the compositional generalization of existing models. Finally, I focus on current practices for evaluation and propose a framework better suited to realistic settings. Current benchmarks for short video understanding have drawbacks, such as employing easy-to-detect distractor answers, not accounting for diversity when depicting the same situation, and not considering realistic settings. I present a novel evaluation format that tackles all these issues and a benchmark that leverages it.

CHAIR: Prof. Rada Mihalcea