Multimodal Systems through a Social Lens: Uncovering and Mitigating Biases

Dorothy Zhao

Department of Computer Science, Stanford University, California, USA

Abstract

Instances of social biases within machine learning systems have been extensively documented, ranging from classifiers exhibiting varying performance based on skin tone to image-text generation models generating stereotypical content. Many efforts have been made to audit AI models and address these identified issues. Multimodal systems present a unique challenge in this space as the sources and pathways for bias propagation become more complicated. In this talk, we will delve into strategies and techniques for both unveiling and mitigating social biases within multimodal systems, utilizing the image-text domain as an illustrative example. More broadly, we want to underscore to researchers and practitioners alike that not only are these biases likely to become more prevalent as AI systems advance in capabilities, but their ramifications will also magnify as these technologies become more common in real-world applications.

1. Short Biography

Dorothy Zhao is a PhD student at Stanford University. Prior to her PhD, she worked as an AI Engineer at Sony Research working on the AI Ethics team. Her research focuses on uncovering, evaluating, and mitigating social biases in AI systems, primarily considering the computer vision and image-text domains. She also is interested in improving dataset creation practices, which spans from collecting novel datasets to understanding challenges that practitioners face in this field.

MUWS'23: 2nd International Workshop on Multimodal Human Understanding for the Web and Social Media, October 22, 2023, Birmingham, UK

Dorothyz@stanford.edu (D. Zhao)