The Role of Commonsense Knowledge in Visual Understanding

Interpreting visual scenes extends beyond just the recognition of observed actions and objects. It involves open world reasoning and constructing deep semantic connections that goes beyond what is directly observed in the video and annotated in the training data. Prior knowledge plays a big role. Current approaches to visual recognition (both images and videos) are highly dependent on deep learning methods, whose success has largely been attributed to the massive amounts of labeled data, whose curation and labeling can be costly and time consuming. Adapting or modifying a model can take almost as much time and energy as creating one from scratch. We show that we can overcome these dependencies by exploiting a diverse set of prior knowledge, including scientific and commonsense knowledge relevant to the problem at hand. We show that the use of commonsense knowledge helps (1) reduce training requirements, (2) interpret scenes beyond observed data and (3) construct inherently explainable interpretations of visual scenes.