Evaluating and Improving Compositionality in Vision and Language

Ranjay Krishna, Associate Professor at University of Washington

Abstract:

Compositionality is a fundamental characteristic of both human vision as well as natural language. It allows us to recognize new scenes and understand new sentences as a composition of previously seen atoms (e.g. objects in images or words in a sentence). Although scholars have spent decades injecting compositional priors into machine learning models, these priors have fallen away with the recent rise of large-scale models trained on internet scale data. In this talk, I will first formalize the notion of compositionality for vision and language by drawing on cognitive science literature. WIth this formalization, we evaluate whether today's best models (including GPT-4V and Gemini) are compositional, uncovering that they perform close to random chance. Next, we will draw on additional priors from neuroscience and cognitive science experiments on human subjects to suggest architectural changes and training algorithms that encourage the emergence of compositionality. Next, we will utilize the same formalism to evaluate generative models, embodied AI, and tool-usage, showcasing that they too are not compositional and demonstrate mechanisms to improve them.