Exploring the mechanisms that enable multimodal reasoning about data visualizations in vision-language models

Alexa Tartaglini
Stanford University, Stanford, California, United States

Christopher Potts
Stanford University, Stanford, California, United States

Judith Fan
Stanford University, Stanford, California, United States

Abstract

Humans can readily integrate visual, linguistic, and numerical information to extract meaning from symbolic displays of information. For instance, answering even a simple question about a data visualization requires connecting tokens of language to visual features in the plot to support quantitative inferences. What are the core computational mechanisms that enable integration across modalities to support such reasoning? Open-source vision-language models (VLMs) might provide a useful testbed for investigating these mechanisms, but doing so requires a high degree of experimental control. To achieve this control, we procedurally generated a large dataset containing pairs of questions and data visualizations that varied along several independent and ecologically important dimensions, including the number of observations and how they were distributed. We identified several open VLMs whose performance was sensitive to this variation, establishing their viability for further exploration of the mechanisms underlying multimodal reasoning.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*