UMEÅ UNIVERSITET

Learning, Reasoning, and Compositional Generalisation in Multimodal Language models

Adam Dahlgren Lindström

Akademisk avhandling

som med vederbörligt tillstånd av Rektor vid Umeå universitet för avläggande av filosofie doktorsexamen framläggs till offentligt försvar i Aula Biologica,Biologihuset, torsdagen den 13 juni, kl. 13:00.

Avhandlingen kommer att försvaras på engelska.

Fakultetsopponent: Professor, Richard Johansson, Chalmers tekniska högskola, Institutionen för Data- och informationsteknik, Göteborg, Sverige.

Institutionen för Datavetenskap

Organization	Document type	Date of publication
Umeå University Department of Computing Science	Doctoral thesis	23 maj 2024

Author
Adam Dahlgren Lindström

Title
Learning, Reasoning, and Compositional Generalisation in Multimodal Language Models

Abstract
Humans learn language and how to interact with the world through our different senses, grounding our language in what we can see, touch, hear, and smell. We call these streams of information different modalities, and our efficient processing and synthesis of the interactions between different modalities is a cornerstone of our intelligence. Therefore, it is important to study how we can build multimodal language models, where machine learning models learn from more than just text. This thesis investigates learning and reasoning in multimodal language models, and their capabilities to compositionally generalise in visual question answering tasks. Compositional generalisation is the process in which we produce and understand novel sentences, by systematically combining words and sentences to uncover the meaning in language, and has proven a challenge for neural networks. The experiments in this thesis compares three neural network-based models, and one neuro-symbolic method, and operationalise language grounding as the ability to reason with relevant functions over object affordances. The results highlight many challenges in multimodal language models, and identify several aspects of how current methods can be improved in the future. The thorough investigation of compositional generalisation suggests that the pretraining of models allow models access to inductive biases that can be useful to solve new tasks. Contrastingly, models trained from scratch show much lower overall performance on the synthetic tasks at hand, but show lower relative generalisation gaps. In the conclusions and outlook, we discuss the implications of these results as well as future research directions.

Keywords
artificial intelligence, natural language processing, multimodal machine learning, reasoning, compositional generalisation, language grounding

Language	ISBN	ISSN	Number of pages
English	print: 978-91-8070-417-5 PDF: 978-91-8070-418-2	0348-0542	204