Towards Faithful Reasoning Using Language Models

Speaker: Antonia Creswell, DeepMind

Abstract:

Language models are showing impressive performance on many natural language tasks, including question-answering. However, language models – like most deep learning models – are black boxes. We cannot be sure how they obtain their answers. Do they reason over relevant knowledge to construct an answer or do they rely on prior knowledge – baked into their weights – which may be biased? An alternative approach is to develop models whose output is a human interpretable, *faithful reasoning* trace leading to an answer. In this talk we will characterise *faithful reasoning* in terms of logically valid reasoning and demonstrate where current *reasoning* models fall short. Following this, we will introduce Selection-Inference, a faithful reasoning model, whose causal structure mirrors the requirements for valid reasoning. We will show that our model not only produces more accurate reasoning traces but also improves final answer accuracy.

Bio:

Antonia Creswell is a Senior Research Scientist at DeepMind, currently working on faithful reasoning. Antonia completed her PhD in Deep Learning and Computer Vision at Imperial College London in 2019, joining DeepMind in the same year. Previously, Antonia obtained an MEng degree in Bioengineering at Imperial College London with a year abroad at the University of California, Davis. Antonia has previously worked as a researcher at Twitter, Cortexica and Knyttan (Unmade).