AI transparency

Contact

Dr. Paweł Widera - pawel.widera@newcastle.ac.uk

Research project

Overview

In recent years, a number of large machine learning models have reached parity with humans on several tasks (e.g. object recognition, speech synthesis, speech/text recognition, etc). However, no one really knows how they work – they are "black boxes", and this is bad for safety. For example, in medical applications where a human life is at stake, we cannot blindly trust that the model will do the right thing. We would like to look inside and confirm that it will not do anything surprising or dangerous.

This project focuses on model transparency and lies at the cross-section of AI alignment/safety and AI explainability research. It aims to develop an ability to audit machine learning models – primarily models based on biomedical data.

Methodology

The working hypothesis on why the large neural network models are able to work so well, and at the same time be so hard to interpret, is related to the concept of compression [1]. In essence, to work well and be easy to interpret, the networks would have to be build from a large number of monosemantic neurons (each representing a single concept). But instead, the networks we build conserve neurons and use a superposition of concepts [2] with polysemantic neurons. Intuitively this could be understood as an attempt at simulating more powerful and larger models that are build from monosemantic neurons [3][4]. Unfortunately, this leads to noise and concept interference and makes our networks hard to interpret, and in consequence, hard to trust when comes to applications with safety requirements, like medicine. However, if we were able to unpack the learned concepts into a monosemantic network, it would be much easier to explain the network behaviour. And this idea lies at the centre of representation/safety engineering, an emerging approach for enhancing AI transparency.

Applicant skills/background

This project requires interest in the topic of AI explainability, experience with applying neural networks (especially to biomedical data), good statistical literacy, professional work habits (team work, planning, openness to challenge), and clear and direct communication skills. A computing science MSc degree in machine learning or applied artificial intelligence will be a bonus. Equivalent practical industry experience is also welcomed.

References

[1] Elhage et al., "*Toy Models of Superposition*", Sep 2022, https://arxiv.org/abs/2209.10652

[2] Chris Olah, "*Distributed Representations: Composition and Superposition*", May 2023, https://transformer-circuits.pub/2023/superposition-composition/index.html

[3] Bills et al., "*Language models can explain neurons in language models*", May 2023, https://openai.com/research/language-models-can-explain-neurons-in-language-models

[4] Zou et al., "*Representation Engineering: A Top-Down Approach to AI Transparency*", Oct 2023, https://arxiv.org/abs/2310.01405