Thesis

Albert Ramus Sidenius Garde (s183969)

February 13, 2024

Contents

Chapter 1

Introduction

- Motivation
  - Large Language Models (LLMs) have shown exceptional performance across a range of tasks, yet their complexity renders their inner workings opaque. This opacity challenges our ability to understand, trust, and safely deploy these models in real-world applications. To rectify this, the overlapping fields of Explainable AI (XAI) and Mechanistic Interpretability (MI) have emerged. The former focuses on developing methods to explain the predictions of LLMs, while the latter focuses on understanding the inner workings of LLMs. Much of the focus in MI has been on understanding the behaviour of individual neurons, but as demonstrated in **?**, the behaviour of individual neurons often doesn't map onto human understandable concepts. **?** shows a possible way forward by providing the features of SAEs (Sparse Autoencoders) as alternative units of interpretability.
- SotA
  - How much should I mention articles that are irrelevant for my own work, but are in the same area?
  -
- Problem statement
  - In this thesis, we attempt to apply the N2G method to the features of SAEs. The goals of this are twofold. Firstly, this is a test of both methods, since if SAEs truly do provide more interpretable features and N2G truly does provide a useful representation of feature behaviour, we would expect this to be reflected when comparing N2G graphs for individual neurons against those for features. Secondly, it is possible that the results could be useful in their own right for understanding models.

Info on what comparisons we expect to make.