Unveiling the Black Box: Causal Inference and Feature Analysis in Fine-Tuned Language Models Using Sparse Autoencoders

Sean Sica, Rini Gupta

4 August 2024

Abstract

3 to 5 sentences Concisely describe your problem, how you solved it, and what you found

1 Introduction

The field of artificial intelligence has witnessed remarkable advancements, with language models like GPT-3 and PaLM demonstrating human-level proficiency in English communication [1]. However, a significant challenge persists: despite creating these powerful systems, we lack a comprehensive understanding of their internal mechanisms. This gap in knowledge not only limits our ability to replicate such systems but also raises concerns about their interpretability and controllability.

Mechanistic interpretability (MI) has emerged as a promising approach to address this challenge. Recent work, particularly by Anthropic, has shown that sparse dictionary learning algorithms, specifically sparse autoencoder (SAE) model architectures, offer a potential path forward in making neural networks more interpretable [?]. Their demonstration of controlling language model outputs using identified features, such as the "Golden Gate Bridge" demo, has inspired further research in this direction.

Our work aims to contribute to this growing field by:

1. Increasing awareness of mechanistic interpretability through accessible feature analysis and causal intervention techniques.
2. Proposing a novel and effective process for evaluating large quantities of SAE features.
3. Testing the hypothesis that fine-tuned models manifest higher quantities of relevant features and sharper, more cohesive features.

The importance of this research lies in its potential to shed light on the fundamental properties of SAE features. We investigate crucial questions such as the relationship between an SAE's training dataset and the resultant features, whether fine-tuning models on specific datasets enhances their interpretability, and if fine-tuned models are easier to steer than their foundational counterparts. These insights are particularly relevant given the widespread use of fine-tuned models in the tech industry.

Our approach addresses a key criticism of current MI work: the potential misinterpretation of SAE features. By employing causal inference techniques, we aim to provide stronger evidence that these features not only are interpretable but also demonstrably represent what we claim they do.

The primary contributions of this paper are:

1. Novel research evaluating sparse autoencoders with respect to fine-tuned models, using randomized controlled trials to better understand fundamental properties of SAE features.
2. Reusable code for practitioners and researchers to easily extract and analyze features from trained sparse autoencoders.
3. A framework for applying causal interventions to steer model behavior conditioned on features of interest, replicating and extending Anthropic's demonstration.

In the following sections, we present our methodology, results, and analysis, providing evidence to substantiate our claims and contribute to the growing body of knowledge in mechanistic interpretability of language models.

2 Background

Present a literature review or discuss related work. This should provide context for your research and show how your work fits into the existing body of knowledge.

3 Methods

Describe your design and implementation in detail. Include any algorithms, models, or experimental setups you used.

4 Results and Discussion

Present your results using tables, plots, and figures as appropriate. Provide a detailed analysis of your findings, comparing them to baselines and relevant literature.

4.1 Subsection 1

Details of your first set of results or analysis.

4.2 Subsection 2

Details of your second set of results or analysis.

5 Conclusion

Summarize your main findings and their implications. Discuss any limitations of your work and potential future directions.

References

[1] John Doe. A sample article. *Journal of Examples*, 2023.

A Appendix

Include any supplementary material, additional data, or extended proofs here.