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 [?]. 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

The foundation of our research lies in Anthropic's work on superposition, polysemanticity, and toy models in neural networks [2, ?, ?, 3]. Their research asserts that neurons in neural networks are polysemantic, responding to mixtures of seemingly unrelated inputs. When inputs are fed into a transformer model, a substantial subset of neurons typically show non-negligible activation. This phenomenon is analogous to observing widespread brain activation in response to specific stimuli during neurological studies.

Anthropic hypothesizes that this behavior results from superposition, where a neural network represents more independent "features" of the data than it has neurons by assigning each feature its own linear combination of neurons. Through experiments with toy models, they demonstrated that sparse dictionary learning can be employed to causally reduce polysemanticity, making inference activations far more sparse and enabling more effective causal inference for neuron labeling.

Scaling the neural network to have more neurons than features is not a viable solution due to incidental polysemanticity. Lecomte et al. [?] suggest that polysemanticity can manifest for multiple reasons, including regularization and neural noise. They propose that incidental polysemanticity might be mitigated by adjusting the learning trajectory without necessarily altering the neural architecture.

In a different approach, researchers at OpenAI have proposed LLM-driven solutions to tackle the interpretability issue [1]. Their method uses prompt engineering to interpret individual neurons. While we approach this method with some skepticism regarding its practicality and reliability, it provides inspiration for our research. We aim to leverage prompt engineering to aid in analyzing and labeling sparse features extracted from sparse autoencoders.

Our research seeks to synthesize Anthropic's approach of decomposing neuron activations into sparse, interpretable features with OpenAI's neuron-centric automated interpretability solution. By combining these methodologies, we aim to enhance the interpretability of fine-tuned language models and provide a more robust framework for understanding their internal mechanisms.

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] Steven Bills, Nick Cammarata, Dan Mossing, Henk Tillman, Leo Gao, Gabriel Goh, Ilya Sutskever, Jan Leike, Jeff Wu, and William Saunders. Language models can explain neurons in language models. `https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html`, 2023.

[2] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. `https://transformer-circuits.pub/2023/monosemantic-features/index.html`.

[3] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022. `https://transformer-circuits.pub/2022/toy_model/index.html`.

[4] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. `https://transformer-circuits.pub/2021/framework/index.html`.

[5] Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024.

A Appendix

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