SPAR Report: Circuit Phenomenology Using Sparse Autoencoders

David Udell [1] **Jackson Kaunismaa** [1] **Hardik Bhatnagar** [1] **Himadri Mandal** [1]

Abstract

Sparse autoencoders provide a means of projecting model activations into a more interpretable vector space. With them, the field of mechanistic interpretability has taken to trying to understand the internals of large language models during training and during inference. In particular, sparse autoencoder dimensions can be naturally composed into circuits–causal directed graphs in which nodes are autoencoder dimensions and edges are causal effects. We looked at the algorithm for recovering these circuits in Marks et al. (2024).

1. Methods

2. Results and Discussion

Contributions

- David Udell:
- Jackson Kaunismaa:
- Hardik Bhatnagar:
- Himadri Mandal:

References

Marks, S., Rager, C., Michaud, E. J., Belinkov, Y., Bau, D., and Mueller, A. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models, 2024. URL https://arxiv.org/abs/2403.19647.

[1]SPAR, Summer 2024 Cohort. Correspondence to: David Udell <udelldavidb@gmail.com>.