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 sparse vector space. With them, the field of mechanistic interpretability has taken to trying to understand the internals of large language models during training and inference. In particular, sparse autoencoder dimensions can be naturally assembled into *circuits* – directed graphs in which nodes are autoencoder dimensions and edges are their causal effects on each other. We looked at an unsupervised algorithm for recovering these circuits in prior work. In reimplementing that algorithm, we isolated a significant bug, with consequences for prior results. Since, we have built out two independent implementations of the circuit discovery algorithm for GPT-2-small (up from Pythia-70m) and are now continuing to work on tuning the graphing hyperparameters involved in graphing upsupervised forward passes.

1. Methods

We built upon prior work due to Marks et al. (2024), specifically focusing on their unsupervised circuit discovery algorithm using attribution patching (rather than integrated gradients, excluding any initial pre-clustering of forward passes based on activation directions). [300-700 WORDS]

2. Results and Discussion

[200-500 WORDS]

Contributions

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

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

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.