Redeeming the Localization Hypothesis

David Reber

Computer Science Department, University of Chicago

Abstract

This project introduces an extension to causal scrubbing with the aim of validating the localization hypothesis in the context of Rank-One Model Editing (ROME) in language models. While previous research has critiqued the efficacy of causal tracing for localizing effective model editing points, this work shifts the focus towards enhancing causal scrubbing methods. By explicitly flipping the input instead of employing noise, the revised approach seeks to address the limitations of traditional causal scrubbing. Theoretical analysis and empirical testing are conducted to explore whether this modification can substantiate the localization hypothesis, offering a novel perspective in the field of mechanistic interpretability.

1 Preface

I was approved to work on the causal scrubbing / ROME paper [Men+22] for my final project directly by David McAllister. The paper has similarly accessibile resources as the other 6 assigned papers, including:

- The original paper[1] and supporting website[2]
- Official Pytorch code: [3] and Colab demos (causal tracing[4], model editing[5])

It's important to note that this project primarily applies to autoregressive models, and so won't use e.g. CIFAR-10. However, a suitable dataset for this project is available, with which GPT-2 XL can be run on Google Colab, making it accessible for interpretability research without extensive data or compute resources.

2 Introduction

In the seminal work by Meng et al. [Men+22], the authors employed *causal tracing* to localize the storage and recall of factual associations in hidden-state activations of certain layers of GPT-2 XL, and then edited these facts via weights using *Rank-One Model Editing* (ROME). However, Hase et al. [Has+23] contends that localization conclusions from causal tracing do not inform which model MLP layer is best to edit using ROME, challenging the *localization hypothesis* and suggesting that increased mechanistic understanding does not necessarily facilitate effective model steering.

This project posits that the failure of causal tracing to localize the best ROME edits is not due to a failure of the localization hypothesis itself, but rather to the inadequacy of causal tracing in providing the promised mechanistic understanding. To support this claim, a

[1] https://arxiv.org/pdf/2202.05262.pdf

[2] https://rome.baulab.info/

[3] https://github.com/kmeng01/rome

[4] https://colab.research.google.com/github/kmeng01/rome/blob/main/notebooks/causal_trace.ipynb

[5] https://colab.research.google.com/github/kmeng01/rome/blob/main/notebooks/rome.ipynb

theoretical analysis is conducted based on the formalism of [RGV23], demonstrating that causal tracing fails to execute sufficiently diverse interventions due to the limited support of noise over flipped base inputs. This theory is then empirically tested to examine if modifying causal scrubbing to explicitly flip the input, as opposed to using noise, is sufficient to rescue the localization hypothesis.

3 Background and Related Work

4 Project Proposal

5 Methodology

6 Data and Resources

7 Expected Outcomes and Evaluation

8 Potential Challenges and Limitations

9 Conclusion

References

[Has+23] P. Hase, M. Bansal, B. Kim, and A. Ghandeharioun. *Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models*. 2023. arXiv: 2301.04213 [cs.LG] (cit. on p. 1).

[Men+22] K. Meng, D. Bau, A. Andonian, and Y. Belinkov. “Locating and editing factual associations in gpt”. *Advances in Neural Information Processing Systems* (2022) (cit. on p. 1).

[RGV23] D. Reber, C. Gârbacea, and V. Veitch. “What’s your Use Case? A Taxonomy of Causal Evaluations of Post-hoc Interpretability”. Accepted to the NeurIPS 2023 Causal Representation Learning workshop. Forthcoming. 2023 (cit. on p. 2).