Extending Causal Tracing to (hopefully) Validate the Localization Hypothesis

David Reber

Computer Science Department, University of Chicago

1 Introduction

In “Locating and Editing Factual Associations in GPT” 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 the MLP weights using *Rank-One Model Editing* (ROME). However, causal tracing recent came under scrutiny by “Does Localization infrom Editing? Surprising Differences in Causality-based Localization vs. Knowledge Editing in Language Models” [Has+23], which 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.

Definition 1. Localization Hypothesis If a ROME edit at a single MLP at layer i and token j is sufficient to restore the uncorrupted label, then intervening on the post-MLP activations at layer i and token j should also flip the label.

Model steering asks ‘how can we change the weights of the network to produce a desired behavior’? If true, the localization hypothesis indicates that we can intervene on activations first (even though these are inherently context-dependent), to narrow down (‘localize’) where we should perform interventions on the MLP weights.

2 Project Proposal

Based on my recent research[1] into the challenge of obtaining a full mechanistic understanding [RGV23], I conjecture 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, I will reimplement the evaluations of [Has+23] with a slight modification to causal tracing, which corrupts the inputs not with noise, but rather with another prompt specifically chosen to flip the label. This extension of causal tracing is inspired by the *interchange interventions* of [Gei+23; Wu+23].

[1] My prior research 1. formalizes the goal of mechantistic interpretability using the language of causality, 2. establishes a taxonomy for how various interpretability methods (including causal tracing) are only addressing portions of mechanistic interpretability, and 3. establishes partial evaluations for these respective methods. **Notably, my past work was only theoretical, so all of the empirical investigations of this project will be new contributions.**

References

[Gei+23] A. Geiger, Z. Wu, C. Potts, T. Icard, and N. D. Goodman. *Finding alignments between interpretable causal variables and distributed neural representations*. 2023. arXiv: 2303.02536 [cs.AI] (cit. on p. 1).

[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. 1).

[Wu+23] Z. Wu, A. Geiger, C. Potts, and N. D. Goodman. *Interpretability at scale: identifying causal mechanisms in alpaca*. 2023. arXiv: 2305.08809 [cs.CL] (cit. on p. 1).