Comparative Analysis of Techniques for Locating Knowledge Editing in Language Models: Integrated Gradients vs. Causal Tracing

Yong Chan Chun[1], Yuna Hur[2*]
[1]Department of Computer Science and Engineering, Korea University, [2]Human-inspired AI Research
{cyc9805, yj72722}@korea.ac.kr

Abstract

This paper compares integrated gradients and causal tracing for knowledge editing in large language models (LLMs). Using GPT2-XL, we evaluate these methods in identifying optimal editing layers. Results on ZsRE dataset show that causal tracing more accurately locates concentrated knowledge, while integrated gradients are useful for editing in layers with diffuse knowledge. These insights guide more precise knowledge editing in LLMs.

Keywords: Knowledge Editing, Causal Mediation Analysis, Integrated Gradient

1. Introduction

Large Language Models (LLMs) encode vast amounts of knowledge but often require updates to correct outdated information. Knowledge editing modifies this knowledge, categorized as either parameter-updating or non-parameter-updating. Among parameter-updating methods, locate-and-edit approaches [1, 2, 3] stand out for efficiently updating knowledge through rank-1 matrices, inspired by linear associative memory in LLMs [4, 5]. Here, the value vectors represent knowledge, while key vectors determine relevance. Locate-and-edit updates only the MLP module's upper-projection matrix.

To reduce computational costs, locate-and-edit identifies layers with relevant knowledge for selective editing [1]. Causal tracing detects indirect neuron effects, while integrated gradients identify neurons responsible for stored knowledge [6, 7]. Our study contrasts layers identified by these methods, assessing their impact on knowledge editing.

2. Attribution Score

The attribution score [7, 8] uses integrated gradients to measure neuron contributions to predictions. For an input prompt x, the probability of correct completion y is:

$$P_x(\hat{w}_i^l) = p(y \mid x, w_i^l = \hat{w}_i^l) \tag{1}$$

where w_i^l denote the original i-th intermediate neuron in the l-th layer, while $\hat{w}_i^l$ denotes modified neuron. The attribution score is calculated as:

$$\text{Attr}(w_i^l) = \bar{w}_i^l \int_0^1 \frac{\partial P_x(\alpha \bar{w}_i^l)}{\partial w_i^l} d\alpha \tag{2}$$

Calculating this continuous integral is computationally intractable, so we approximate the attribution score using a Riemann sum:

$$\text{Attr}(w_i^l) \approx \frac{\bar{w}_i^l}{n} \sum_{k=1}^{n} \frac{\partial P_x\left(\frac{k}{n} \bar{w}_i^l\right)}{\partial w_i^l} \tag{3}$$

3. Comparison with Causal Tracing

Using GPT2-XL [9] and factual statements from [1], we compute attribution scores to locate key layers. Integrated gradients identify earlier layers, while causal tracing targets intermediate layers, as shown in Table 1.

4. Experiments

We apply ROME [1] and MEMIT [2] to layers identified by attribution scores and causal tracing. For attribution, ROME edits layer 1, and MEMIT edits layers 1–5. For causal tracing, ROME and MEMIT edit layers 17 and 13–17, respectively.

4.1 Implementation and Results

Editing on GPT2-XL with the ZsRE [10] dataset, we assess reliability, locality, portability, fluency, and overall

*Corresponding author

Table 1. Comparison of ROME and MEMIT Performance on the ZsRE Dataset.

	ROME					MEMIT				
	Reliability	Locality	Portability	Fluency	Score	Reliability	Locality	Portability	Fluency	Score
Attribution score	65.58	56.91	27.23	493.46	43.14	48.31	**87.98**	31.65	**596.52**	47.13
Causal tracing	**99.88**	**67.83**	**35.04**	**577.19**	**56.29**	**64.35**	81.89	**32.12**	593.56	**50.95**

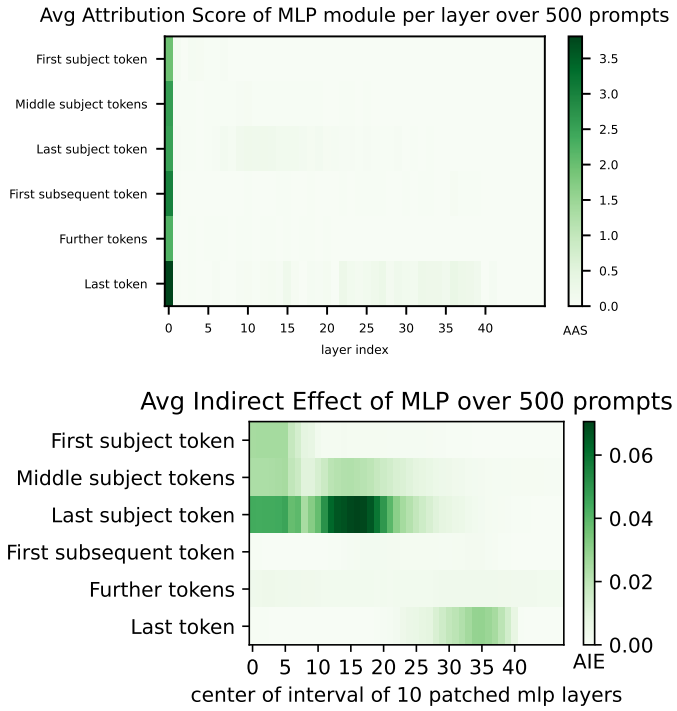

Figure 1. Comparison between attribution scores (top) and causal tracing methods (bottom) for all input prompts. Attribution scores are multiplied by 10,000 for clearer visualization.

score. Results in Tables 1 show causal tracing generally outperforms integrated gradients in reliability, portability, and score, while integrated gradients better preserve unrelated knowledge, improving locality and fluency.

5. Conclusion

We compare integrated gradients and causal tracing for identifying optimal LLM layers for knowledge editing. Causal tracing excels at locating concentrated knowledge, while integrated gradients better preserve unrelated content, advancing context-aware editing in LLMs.

Acknowledgements

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT). (RS-2024-00398115, Research on the reliability and coherence of outcomes produced by Generative AI). This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2021R1A6A1)

Reference

[1] K. Meng, D. Bau, A. Andonian, and Y. Belinkov, "Locating and editing factual associations in gpt," 2023.

[2] K. Meng, A. S. Sharma, A. Andonian, Y. Belinkov, and D. Bau, "Mass-editing memory in a transformer," 2023.

[3] X. Li, S. Li, S. Song, J. Yang, J. Ma, and J. Yu, "Pmet: Precise model editing in a transformer," 2024.

[4] T. Kohonen and M. Ruohonen, "Representation of associated data by matrix operators," *IEEE Transactions on Computers*, Vol. C-22, No. 7, pp. 701–702, 1973.

[5] J. A. Anderson, "A simple neural network generating an interactive memory," *Mathematical Biosciences*, Vol. 14, No. 3, pp. 197–220, 1972.

[6] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," 2017.

[7] D. Dai, L. Dong, Y. Hao, Z. Sui, B. Chang, and F. Wei, "Knowledge neurons in pretrained transformers," 2022.

[8] Y. Hao, L. Dong, F. Wei, and K. Xu, "Self-attention attribution: Interpreting information interactions inside transformer," 2021.

[9] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," 2019.

[10] O. Levy, M. Seo, E. Choi, and L. Zettlemoyer, "Zero-shot relation extraction via reading comprehension," 2017.