<ICML2026 Mechanistic Intepretability Workshop Virtual Poster>

Dual Path Attribution: Efficient Attribution for SwiGLU-Transformers through Layer-Wise Target Propagation

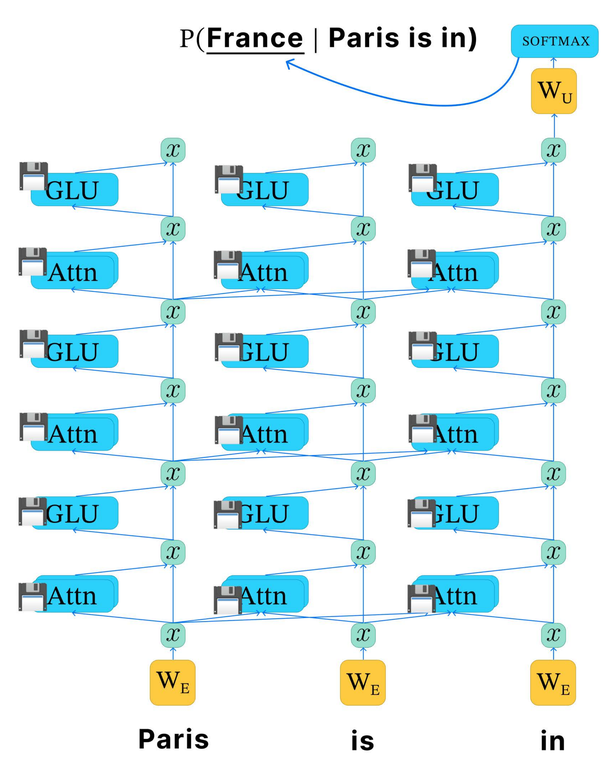
Lasse Jantsch*[1], Dong-Jae Koh*[1], Seonghyeon Lee†[1], Young-Kyoon Suh†[1]

*Equal Contributions †Corresponding author

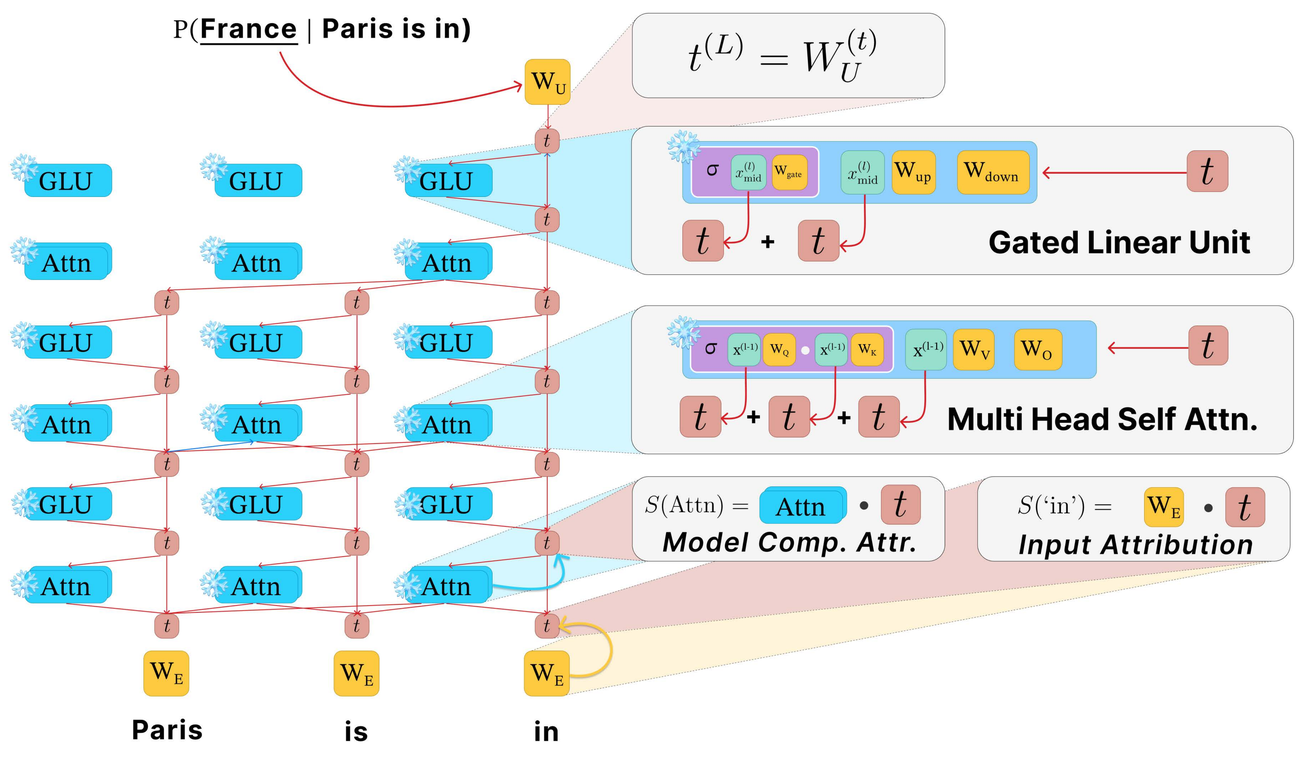
[1]School of Computer Science and Engineering, Kyungpook National University

{lassejantsch, djkoh, sh0416, yksuh}@knu.ac.kr

Dual Path Attribution achieves **dense component attribution at O(1) cost** per component by back-propagating targets to each residual position.



1. Cache activations on forward pass.



2. Propagate target through frozen modules.

Background

A wide range of attribution methods has been proposed gradient-based, attention-based, and decomposition-based. Each, however, remains bound by the same ***trade-off between faithfulness and computational cost.***

Our Approach

Forward pass

Cache activation values and calculate target logit.

Backward pass

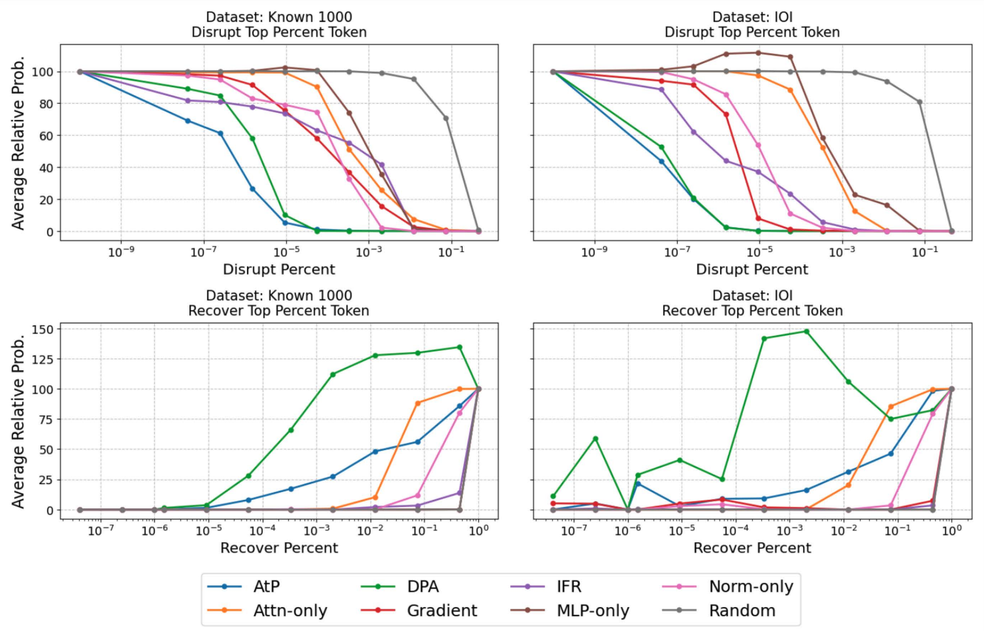
Back-propagate target logit to each residual position on frozen forward pass.

Attribution

Component attribution with O(1) cost by calculating the dot product.

Component Attribution Results

Methods	Known 1000			IOI		
	dis. ↓	rec. ↑	total ↑	dis. ↓	rec. ↑	total ↑
Reference						
Random	19.77	27.55	7.78	21.85	27.54	5.69
Baselines						
Attn-only	0.66	93.43†	92.77†	0.15	**93.07**	**92.92**
MLP-only	0.37	27.58	27.21	0.87	27.53	26.66
Norm-only	0.06†	67.25	67.19	0.01†	65.02	65.01
Gradient-based						
Gradient	0.28	27.62	27.34	**0.00**	30.95	30.95
AtP	**0.00**	81.39	81.39	**0.00**	84.39	84.39
Contextual Mixing-based						
IFR	0.32	34.72	34.40	0.02	29.15	29.13
Proposed Method						
Dual Path Attribution (DPA)	**0.00**	**123.50**	**123.50**	**0.00**	86.87†	86.87†



Conclusion

- **State-of-the-art faithfulness** for a range of models and datasets.
- **O(1) cost per component** lead to a method scalable to larger models and longer input sequences.

Scan to **download full paper**

The 43rd International Conference on Machine Learning
Mechanistic Intepretability Workshop