**Mechanistic Interpretability** aims to uncover causal relationships and precise computations transforming inputs into outputs. Mechanistic Interpretability methods can be categorized into the following three types:

**(1) Intrinsic interpretability methods** encourage sparsity, modularity and monosemanticity in architectural choices and training process of neural networks to reverse engineering without sacrificing performance.

For example, [cite 3] proposes a method to achieve monosemanticity in language models through using a bilinear layer instead of a linear layer. [cite 1] and [cite 2] have demonstrated that brain-inspired modular, a spatial regularization regime, can make RNNs exhibit brain-like anatomical modularity.

**(2) Developmental interpretability methods** offers insights into the emergence of structures and behaviors over time, compared to static analyses. It can help researchers discover fundamental principles, which may not be apparent in static analyses.

For example, singular Learning Theory (SLT) is a rigorous framework to explain the behavior and generalization of overparameterized models. [cite 4] quantifies model complexity, and offers insights into learning phase transitions.

**(3) Post-hoc interpretability methods** explain the internal computational logic and identify human-understandable representations or abstractions after training. Within post-hoc methods, we distinguish between local/narrow approaches and global/comprehensive approaches.

For example, circuit-style mechanistic interpretability [cite 5, cite 6] aims to explain neural networks by reverse engineering the underlying mechanisms at the level of individual neurons or subgraphs, aims to discover subnetworks (circuits) responsible for specific capabilities.

Feature attribution methods [cite 7, cite 8] usually aim to provide attribution scores for each individual input unit. However, most feature attribution methods can only estimate the importance score for each input variable, they cannot theoretically guarantee the explained importance score reflect the real inference logic of the DNN.

**Distinctive contribution of interaction theory.** The interaction-based method explains the trained model's internal inference logic as a small number of salient interactions in a post-hoc manner. **Unlike previous studies, theorems of the universal matching property and sparsity property first proves that the detailed inference logic of a DNN can be faithfully explained by a small number of symbolic interactions.** Thanks to the theoretically guaranteed faithfulness, the interaction has been used to explain the generalization power [cite 9], the adversarial robustness [cite 10], the learning dynamics [cite 11] of a DNN, and are used to explain the internal mechanisms of 14 adversarial-transferability-boosting methods [cite 12], and 12 attribution methods [cite 13].

In our study, we first extend AND interactions to OR interactions, which enables us to simultaneously explain both AND relationship and OR relationship encoded by the DNN. Our technology let the research jump out from the low-level neuron circuits or local causal structures, and directly examine the AND-OR interaction concepts/logic encoded by the DNN. In comparison, other concept-level/logic-level explanation method usually rely on heuristic probes and interventions in a coarse manner without touching the very detailed logic or theoretically guaranteed numerical accuracy.


[cite 1] Ziming Liu, Eric Gan, and Max Tegmark. Seeing is believing: Brain-inspired modular training for mechanistic interpretability. Entropy. (2023)

[cite 2] Ziming Liu, Mikail Khona, Ila R. Fiete, and Max Tegmark. Growing brains: Co-emergence of anatomical and functional modularity in recurrent neural networks. CoRR. (2023)

[cite 3] Lee Sharkey. A technical note on bilinear layers for interpretability. CoRR. (2023)

[cite 4] Edmund Lau, Daniel Murfet, and Susan Wei. Quantifying degeneracy in singular models via the learning coefficient. CoRR, (2023)

[cite 5] Tom Lieberum, Matthew Rahtz, János Kramár, Neel Nanda, Geoffrey Irving, Rohin Shah, and Vladimir Mikulik. Does circuit analysis interpretability scale? evidence from multiple choice capabilities in chinchilla. CoRR. (2023)

[cite 6] Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. ICLR. (2023)

[cite 7] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. ICCV. (2017)

[cite 8] Mukund Sundararajan, Ankur Taly, Qiqi Yan. Axiomatic attribution for deep networks. ICML. (2017)

[cite 9] Huilin Zhou, Hao Zhang, Huiqi Deng, Dongrui Liu, Wen Shen, Shih-Han Chan, Quanshi Zhang. Explaining generalization power of a dnn using interactive concepts. AAAI. (2024)

[cite 10] Jie Ren, Die Zhang, Yisen Wang, Lu Chen, Zhanpeng Zhou, Yiting Chen, Xu Cheng, Xin Wang, Meng Zhou, Jie Shi, Quanshi Zhang. Towards a unified game-theoretic view of adversarial perturbations and robustness. NIPS. (2021)

[cite 11] Qihan Ren, Junpeng Zhang, Yang Xu, Yue Xin, Dongrui Liu, Quanshi Zhang. Towards the dynamics of a DNN learning symbolic interactions. NIPS. (2024)

[cite 12] Xin Wang, Jie Ren, Shuyun Lin, Xiangming Zhu, Yisen Wang, Quanshi Zhang. A Unified Approach to Interpreting and Boosting Adversarial Transferability. ICLR. (2021)

[cite 13] Huiqi Deng, Na Zou, Mengnan Du, Weifu Chen, Guocan Feng, Ziwei Yang, Zheyang Li, Quanshi Zhang. Unifying fourteen post-hoc attribution methods with taylor interactions. PAMI. (2024)