High-order Component Attribution via Kolmogorov-Arnold Networks

Anonymous Author(s)
Affiliation
Address
email

Abstract

Component attribution methods provide insight into how parts of deep learning models, such as convolutional filters and attention heads, influence model predictions. Despite their successes, existing attribution approaches typically assume component effects are additive and independent, neglecting complex interactions among components. Capturing these relations between components is crucial for a better mechanistic understanding of these models. In this work, we improve component attribution (COAR) by replacing the linear counterfactual estimator with a Kolmogorov–Arnold Network (KAN) surrogate fitted to example-wise perturbation–response data. Then, a symbolic approximation of the learned KAN lets us compute mixed partial derivatives that captures and makes explicit high-order component interactions that linear methods are missing. These symbolic expressions facilitate future integration with formal verification methods, enabling richer counterfactual analyses of internal model behavior. Preliminary results on standard image classification models demonstrate that our approach improves the accuracy of predicted counterfactuals and enable extraction of higher-order component interactions compared to linear attribution methods.

1 Introduction

Advances in deep learning generate continuous performance improvements across various tasks, including image classification, language modeling, and audio processing [1–3]. However, the growing complexity of deep models often obscures the precise role of individual components such as transformer blocks, residual blocks, or convolutional layers play in generating specific predictions [4, 5]. Component attribution methods, which quantify the effect of ablating or perturbing these components on model outputs, have become essential tools for interpretability and targeted interventions [6].

Existing attribution methods, including component attribution via regression (COAR) [5], typically assume additive independence among component effects. Although these methods are efficient and insightful, their linear assumptions limit their ability to capture complex, nonlinear interactions between components. Addressing these higher-order interactions explicitly is critical to advancing our mechanistic understanding of deep neural networks.

In this paper, we propose a novel nonlinear component attribution framework based on Kolmogorov–Arnold Networks (KANs) [7, 8]. Our method builds upon the formalization of component modeling introduced by Shah et al. [5], replacing the linear attribution approach with a flexible nonlinear component model. Specifically, we construct perturbation-response datasets from randomized, continuous multicomponent interventions. We then train a KAN to approximate these responses and use symbolic regression [9] to derive a closed-form symbolic expression. Higher-order mixed partial

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

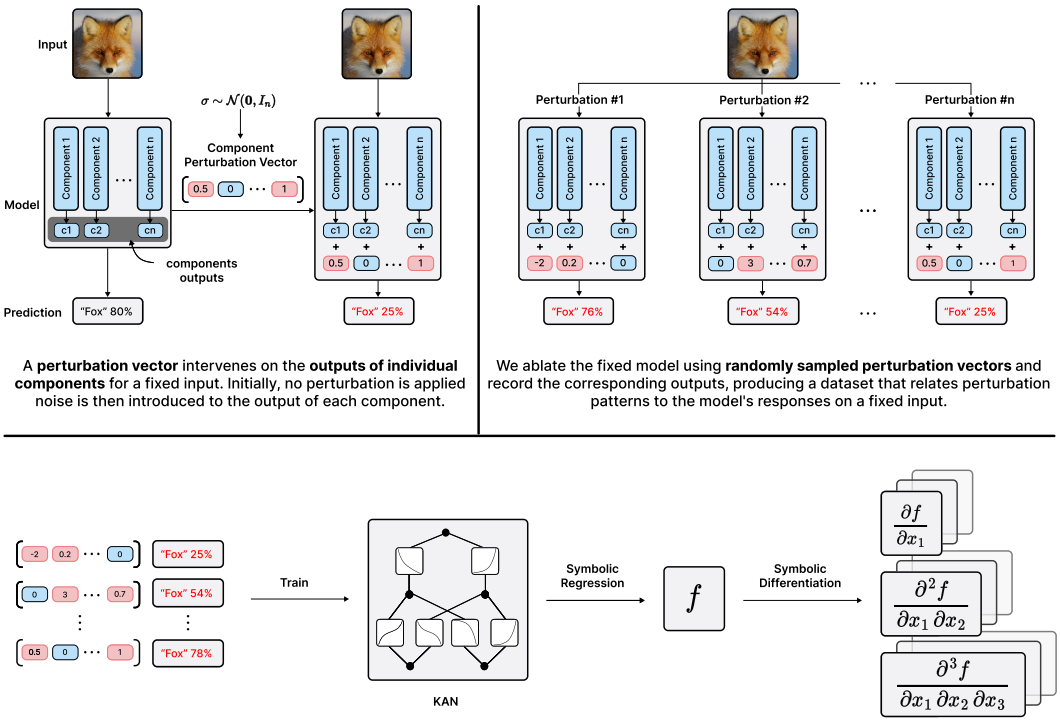

Figure 1: Overview of the proposed approach.

derivatives of this symbolic function directly quantify interactions among components. Figure 1 illustrates our overall approach.

Our main contributions are:

1. Formalizing higher-order component attribution explicitly within the established component modeling framework [5].
2. Introducing a three-step methodological approach that:
   (a) Fits nonlinear KAN component models to perturbation-response datasets.
   (b) Makes a symbolic approximation of the learned component model and analytically computes their derivatives.
   (c) Quantifies higher-order component interactions via mixed partial derivatives.
3. Providing empirical validation that nonlinear modeling significantly enhances the accuracy of counterfactual predictions compared to linear approaches.

2 High-order Component Attribution via KANs

Setup We build upon the component modeling framework introduced by Shah et al. [5]. Given a trained model M composed of m components $\mathcal{C} = \{c_1, \ldots, c_m\}$ and a fixed input $\boldsymbol{z}$, we define the scalar output function $f_M(\mathbf{z}, \boldsymbol{\sigma})$ as the model's prediction when applying an additive gating mask $\boldsymbol{\sigma} \sim \mathcal{N}_m(\mathbf{0}, \mathbf{I}_m)$ to its components. Under this definition, the mask $\mathbf{0}$ represents the unperturbed model, while any deviation from $\mathbf{0}$ corresponds to partial ablation or amplification of component outputs. We formalize this notion through the centered counterfactual response:

$$\Delta f_M(\mathbf{z}, \boldsymbol{\sigma}) = f_M(\mathbf{z}, \boldsymbol{\sigma}) - f_M(\mathbf{z}, \mathbf{0}) \tag{1}$$

ensuring $\Delta f_M(\boldsymbol{z}, \mathbf{0}) = 0$ by construction.

Perturbation-response Dataset To build a perturbation–response dataset D for each input $\mathbf{z}$, we draw N random masks $\boldsymbol{\sigma}^{(i)}$ from a multivariate normal distribution $\mathcal{N}_m(\mathbf{0}, \mathbf{I}_m)$ (Algorithm 1). For every mask, we record the centered output $y^{(i)} = \Delta f_M(\mathbf{z}, \boldsymbol{\sigma}^{(i)})$. The resulting dataset captures local nonlinear dependencies among components in the neighborhood of the intact configuration.

KAN as a component model We train a per-example KAN component model $g_{\boldsymbol{z}} : \mathbb{R}^m \to \mathbb{R}$ using the dataset D. The component model aims to approximate the nonlinear mapping from component perturbations to changes in model outputs by minimizing:

$$\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \left(g_{\boldsymbol{z}}(\boldsymbol{\sigma}^{(i)}; \theta) - y^{(i)} \right)^2 \tag{2}$$

Symbolic Approximation and Interaction Scores After training, we symbolically approximate each univariate spline function within the trained KAN component model g_z via symbolic regression [9], yielding a closed-form symbolic component model $\hat{g}_{\mathbf{z}}$ (Algorithm 2).

To quantify interactions of arbitrary order, we compute mixed partial derivatives of $\hat{g}_{\mathbf{z}}$ at the intact configuration $\boldsymbol{\sigma} = \mathbf{0}$. For any subset of component indices $S \subseteq 1, \ldots, m$ of size $r = |S|$, we define the local r-way interaction coefficient as:

$$L_S^{(r)}(\mathbf{z}) = \left. \frac{\partial^r \hat{g}_z(\boldsymbol{\sigma})}{\prod_{j \in S} \partial \sigma_j} \right|_{\boldsymbol{\sigma} = \mathbf{0}} \tag{3}$$

Algorithm 3 systematically computes these coefficients up to a specified order k, providing hierarchical insights into component effects: first-order terms quantify independent effects, while higher-order terms reveal complex joint interactions. When nonlinear interactions are negligible, our method naturally reduces to standard linear component attribution [5]

2.1 Experimental Setup

We follow the experimental framework proposed by Shah et al. [5], adapted to our specific level of granularity. Rather than examining individual neurons, we focus on residual blocks in ResNet architectures [10] and encoder layers in Vision Transformers (ViTs) [11] as individual components. This granularity is selected due to the intrinsic limitations of Kolmogorov–Arnold Networks (KANs), which are currently unable to efficiently handle high-dimensional inputs, thereby constituting a limitation of our approach. We assess our methodology on three widely used image classification setups: ResNet-18 trained on CIFAR-10 [12], ResNet-50 trained on ImageNet [13], and ViT-B/16 also trained on ImageNet. For each model-dataset combination, we generate localized perturbation-response datasets by sampling multicomponent perturbations, following Algorithm 1.

Baselines and Evaluation Metrics We benchmark our proposed approach against the established linear attribution baseline:

- **COAR (Linear Attribution)** [5]: This baseline applies a linear model to perturbation-response data.

In line with previous research [5], we quantify attribution accuracy using two metrics: the Pearson correlation coefficient and the mean squared error, computed between predicted and actual responses on held-out perturbations.

Our experimental results, summarized in Table 1, show that the proposed KAN-based approach consistently outperforms the linear COAR baseline across all evaluated tasks. Specifically, KAN achieves higher Pearson correlations and lower mean squared errors, indicating a clear advantage in modeling nonlinearity within the component attribution framework.

3 Conclusions and Future Work

In this work, we introduced a novel nonlinear component attribution framework based on Kolmogorov–Arnold Networks (KANs), capable of explicitly capturing higher-order component interactions that traditional linear attribution methods overlook. Our experiments demonstrated that

Table 1: Comparison of attribution accuracy between predicted and observed counterfactual responses

	ResNet-18 / CIFAR-10		ResNet-50 / ImageNet		ViT-B/16 / ImageNet	
Method	Pearson ↑	MSE ↓	Pearson ↑	MSE ↓	Pearson ↑	MSE ↓
COAR (Linear)	0.62 ± 0.05	0.82 ± 0.07	0.54 ± 0.06	0.95 ± 0.08	0.48 ± 0.07	1.05 ± 0.09
KAN (Ours)	$\mathbf{0.76 \pm 0.03}$	$\mathbf{0.53 \pm 0.05}$	$\mathbf{0.70 \pm 0.04}$	$\mathbf{0.60 \pm 0.06}$	$\mathbf{0.66 \pm 0.05}$	$\mathbf{0.68 \pm 0.07}$

employing KAN-based component models significantly improves the accuracy of predicted counterfactual responses compared to linear baselines. However, the current methodology is limited by the intrinsic dimensionality constraints inherent to KANs, restricting their scalability to high-dimensional input spaces. Future research should focus on addressing these dimensionality challenges to broaden the applicability and scalability of the proposed method. Additionally, it will be valuable to explore meaningful use cases of higher-order interactions in realistic scenarios. While our current results already show promising improvements in attribution accuracy, further extensions could illustrate practical benefits of higher-order attributions enabled by our proposed methodology, potentially leading to deeper insights into complex model behaviors.

4 Algorithms

Algorithm 1 Perturbation–Response Dataset

procedure GENERATEDATASET(example $\mathbf{z}$, model M with components $\mathcal{C}$ (size m), sample size N)
  $D \leftarrow []$ ▷ init dataset
  for $i \in \{1, \dots, N\}$ **do** ▷ N samples
    Sample $\boldsymbol{\sigma}^{(i)} \sim \mathcal{N}_m(\mathbf{0}, \mathbf{I}_m)$ ▷ multicomponent perturbation
    $\Delta f_M(\mathbf{z}, \boldsymbol{\sigma}^{(i)}) = f_M(\mathbf{z}, \boldsymbol{\sigma}^{(i)}) - f_M(\mathbf{z}, \mathbf{0})$
    $y^{(i)} \leftarrow \Delta f_M(\mathbf{z}, \boldsymbol{\sigma}^{(i)})$ ▷ model output
    $D \leftarrow D + [(\boldsymbol{\sigma}^{(i)}, y^{(i)})]$ ▷ append pair
  end for
  return D ▷ dataset for surrogate
end procedure

Algorithm 2 Symbolic KAN Surrogate

procedure SYMBOLICSURROGATE(dataset D)
  Fit KAN $g_{\mathbf{z}}$ on D ▷ train surrogate
  $\mathcal{S} \leftarrow []$ ▷ init symbol list
  for edge e in $g_{\mathbf{z}}$ **do** ▷ per-edge univariate
    $\hat{\phi}_e \leftarrow \mathrm{SymbolicRegression}(\phi_e)$ ▷ closed-form fit
    $\mathcal{S} \leftarrow \mathcal{S} + [(e, \hat{\phi}_e)]$ ▷ collect
  end for
  Replace $\phi_e \leftarrow \hat{\phi}_e$ in $g_{\mathbf{z}}$ ▷ compose symbolic surrogate
  return $\hat{g}_{\mathbf{z}}, \mathcal{S}$ ▷ outputs
end procedure

Algorithm 3 Local Interaction Coefficients

procedure LOCALINTERACTIONS(symbolic surrogate $\hat{g}_{\mathbf{z}}$, components $\mathcal{C}$ (size m), max order k)
  for $r \in \{1, \dots, k\}$ **do** ▷ interaction order
    for index subset $S \subseteq \{1, \dots, m\}$ with $|S| = r$ **do** ▷ choose indices
      $L_S^{(r)} \leftarrow \left. \dfrac{\partial^r \hat{g}_{\mathbf{z}}(\boldsymbol{\sigma})}{\prod_{j \in S} \partial \sigma_j} \right|_{\boldsymbol{\sigma} = \mathbf{0}}$ ▷ r-way interaction at baseline
    end for
  end for
  return $\{L^{(r)}\}_{r=1}^{k}$ ▷ local interaction coefficients
end procedure

References

[1] Jie Huang and Kevin Chen-Chuan Chang. Towards Reasoning in Large Language Models: A Survey. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.67. URL https://aclanthology.org/2023.findings-acl.67/.

[2] Yutong Zhou and Nobutaka Shimada. Vision + Language Applications: A Survey. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 826–842, Vancouver, BC, Canada, June 2023. IEEE. ISBN 979-8-3503-0249-3. doi: 10.1109/CVPRW59228.2023.00090. URL https://ieeexplore.ieee.org/document/10208464/.

[3] Siddique Latif, Moazzam Shoukat, Fahad Shamshad, Muhammad Usama, Yi Ren, Heriberto Cuayáhuitl, Wenwu Wang, Xulong Zhang, Roberto Togneri, Erik Cambria, and Björn W. Schuller. Sparks of Large Audio Models: A Survey and Outlook, September 2023. URL http://arxiv.org/abs/2308.12792. arXiv:2308.12792 [cs].

[4] Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, Stella Biderman, Adria Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, Eric J. Michaud, Stephen Casper, Max Tegmark, William Saunders, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Tom McGrath. Open Problems in Mechanistic Interpretability, January 2025. URL http://arxiv.org/abs/2501.16496. arXiv:2501.16496 [cs].

[5] Harshay Shah, Andrew Ilyas, and Aleksander Madry. Decomposing and Editing Predictions by Modeling Model Computation, April 2024. URL http://arxiv.org/abs/2404.11534. arXiv:2404.11534 [cs].

[6] Shichang Zhang, Tessa Han, Usha Bhalla, and Himabindu Lakkaraju. Towards Unified Attribution in Explainable AI, Data-Centric AI, and Mechanistic Interpretability, May 2025. URL http://arxiv.org/abs/2501.18887. arXiv:2501.18887 [cs].

[7] Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y. Hou, and Max Tegmark. KAN: Kolmogorov-Arnold Networks, April 2024. URL http://arxiv.org/abs/2404.19756. arXiv:2404.19756 [cs] version: 1.

[8] Ziming Liu, Pingchuan Ma, Yixuan Wang, Wojciech Matusik, and Max Tegmark. KAN 2.0: Kolmogorov-Arnold Networks Meet Science, August 2024. URL http://arxiv.org/abs/2408.10205. arXiv:2408.10205 [cs].

[9] Miles Cranmer. Interpretable Machine Learning for Science with PySR and SymbolicRegression.jl, May 2023. URL http://arxiv.org/abs/2305.01582. arXiv:2305.01582 [astro-ph].

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, USA, June 2016. IEEE. ISBN 978-1-4673-8851-1. doi: 10.1109/CVPR.2016.90. URL http://ieeexplore.ieee.org/document/7780459/.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021. URL http://arxiv.org/abs/2010.11929. arXiv:2010.11929 [cs].

[12] Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database.