Feature Attribution with Necessity and Sufficiency via Dual-stage Perturbation Test for Causal Explanation

Xuexin Chen[1], Ruichu Cai[1,5], Zhengting Huang[1], Yuxuan Zhu[1], Julien Horwood[2], Zhifeng Hao[3], Zijian Li[4], José Miguel Hernández-Lobato[2]

[1]*Guangdong University of Technology, Guangzhou, China* [2]*University of Cambridge, Cambridge, UK,* [3]*Shantou University, Shantou, China*
[4]*Mohamed bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE* [5]*Pazhou Laboratory (Huangpu), Guangzhou, China*

Email: cairuichu@gmail.com

Motivation

- To explain the ML model, feature attribution methods assign weights to input features through a perturbation test, i.e., comparing the difference in prediction under different perturbations.
- However, this perturbation test may not accurately distinguish the contributions of different features to the prediction when their changes in prediction are similar after perturbation.

Causal Model for Feature Attribution

We develop a principled causal framework to model the perturbation test in feature attribution.

- First, we define the neighborhood of d-dimensional target input $\mathbf{x}^t$ to be explained as the distribution of $\mathbf{X}$ on a dimension subset $\mathbf{S} \subseteq \{1, \dots, d\}$:

$$\widetilde{\mathbf{X}} \sim P(\mathbf{X} \mid \|\mathbf{X_s} - \mathbf{x_s}^t\|_p \leq b)$$

- Second, draw a sample in the neighborhood of $\mathbf{x}^t$.
- Finally, use a perturbation function g to introduce perturbations by replacing features on $\mathbf{S}$ with the baseline value $\mathbf{x}'_s$, and use the model f to generate a new prediction $\mathbf{Y}$:

$$\mathbf{Y} = f(g(\widetilde{\mathbf{X}}, \mathbf{S}, \mathbf{x}'))$$

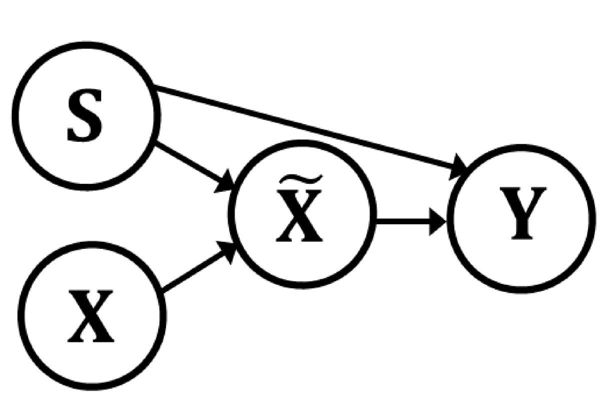

Fig. 1 Causal diagram of standard perturbation test in feature attribution.

Feature Attribution as a Problem of PNS Measurement

$A_{s,b}$: An event of perturbation on the subset $\mathbf{s}$ of an input sampled from $P(\mathbf{X} \mid \|\mathbf{X_s} - \mathbf{x_s}^t\|_p \leq b)$.

$B_{s,c}$: An event representing the change in the original prediction $f(\mathbf{x})$ relative to the prediction $f(\mathbf{x}')$ of perturbed $\mathbf{x}'$, expressed as $|f(\mathbf{x}') - f(\mathbf{x})| > c$.

PN: $PN = P(\bar{B}_{\mathbf{s},c_{A_{s,b}}} \mid A_{\mathbf{s},b}, B_{\mathbf{s},c})$

PS: $PS = P(B_{\mathbf{s},c_{\bar{A}_{s,b}}} \mid \bar{A}_{\mathbf{s},b}, \bar{B}_{\mathbf{s},c})$

PNS

$$PNS = PN \cdot P(A_{\mathbf{s},b}, B_{\mathbf{s},c}) + PS \cdot P(\bar{A}_{\mathbf{s},b}, \bar{B}_{\mathbf{s},c})$$

- To evaluate the importance w_s of the target input $\mathbf{x}^t$ on the dimension subset s, we aim for w_s finding a neighborhood for $\mathbf{x}^t$ where perturbing samples within this neighborhood on s have the highest probability of being necessary and sufficient causes for prediction change, and take this probability as w_s. Mathematically,

Definition (**Necessary and Sufficient Attribution**) Necessary and Sufficient Attribution of the input $\mathbf{x}^t$ on the dimension subset $\mathbf{s}$ is defined as

$$w_{\mathbf{s}} := \max_{b,c} PN \cdot P(A_{\mathbf{s},b}, B_{\mathbf{s},c}) + PS \cdot P(\bar{A}_{\mathbf{s},b}, \bar{B}_{\mathbf{s},c})$$

Algorithm: Feature Attribution with Necessity and Sufficiency (FANS)

We implement our FANS to compute the Necessary and Sufficient Attribution of the input $\mathbf{x}^t$ on the dimension subset $\mathbf{s}$.

- Calculation of b, c: Boundary b is calculated using the Scott rule to ensure a small and nearly uniformly dense neighborhood. Threshold c is determined by the maximum variance of the model's predictions under low-intensity noise simulated by a Gaussian distribution.
- Calculation of PS: Design a dual-stage perturbation test to estimate PS.

  1) Factual stage: Draw inputs from a distribution conditional on the fact that predictions remain unchanged after applying perturbation to the features of $\widetilde{\mathbf{X}}$ on $\bar{\mathbf{s}}$.

  2) Since the conditional distributions are complex, FANS employs the Sampling-Importance-Resampling to approximate these distributions using observed samples.

  3) Intervention stage: Apply perturbation to the features of $\widetilde{\mathbf{X}}$ on $\mathbf{s}$ and calculate the proportion of prediction changes

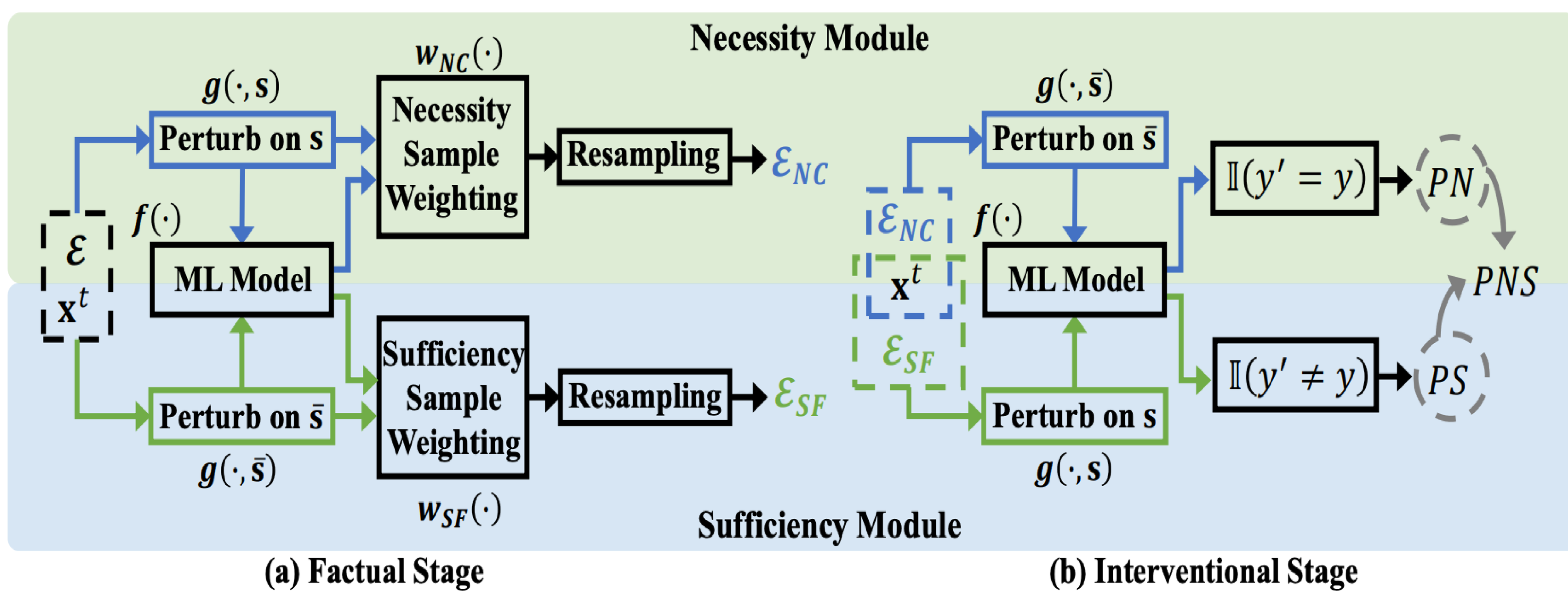

Fig. 2 Architecture of FANS, which takes the sample $\mathbf{x}^t$ to be explained and the samples $\mathcal{E} \overset{iid}{\sim} P(\mathbf{X})$ as inputs, throughout the necessity and sufficiency modules to output PN and PS, and finally combine PN, PS into PNS.

Experiment

- Performance of FANS. Table below shows FANS consistently outperforms all baselines in faithfulness, sparsity, and robustness across various image datasets.

Method	MNIST INF↓	MNIST IR↑	MNIST SPA↑	MNIST MS↓	Fashion-MNIST INF↓	Fashion-MNIST IR↑	Fashion-MNIST SPA↑	Fashion-MNIST MS↓	CIFAR10 INF↓	CIFAR10 IR↑	CIFAR10 SPA↑	CIFAR10 MS↓
Saliency	3.8×10^4	64.3	0.658	0.623	1.8×10^6	25.5	0.558	0.753	1.2×10^8	54.1	0.492	0.736
IG	1.7×10^3	73.3	0.918	0.683	1.7×10^4	60.8	0.612	0.806	1.5×10^5	63.5	0.631	0.966
DeepLift	2.2×10^3	73.3	0.918	0.679	8.2×10^4	59.8	0.610	0.797	1.9×10^5	62.7	0.631	0.959
IDGI	2.0×10^3	64.7	0.837	0.578	2.6×10^4	58.2	0.593	0.781	2.3×10^4	19.5	0.632	0.854
GradShap	2.2×10^3	73.3	0.918	0.673	2.5×10^4	59.2	0.614	0.874	2.3×10^5	56.5	0.630	1.000
LIME	6.6×10^5	67.9	0.808	0.899	2.5×10^8	28.6	0.533	0.884	1.1×10^9	2.1	0.512	1.032
Occlusion	5.7×10^5	69.2	0.802	0.538	2.8×10^7	58.4	0.505	0.660	2.1×10^8	51.4	0.507	0.857
FeatAblation	1.7×10^3	72.8	0.917	0.669	5.5×10^4	45.4	0.572	0.791	1.1×10^6	36.8	0.619	0.983
MP	8.5×10^5	70.3	0.421	0.904	2.9×10^7	20.1	0.227	**0.453**	5.1×10^8	16.3	0.476	0.887
CIMI	1.7×10^4	12.9	0.901	0.548	2.6×10^4	54.8	0.589	0.827	6.0×10^6	21.1	0.589	0.615
FANS	$\mathbf{9.0 \times 10^2}$	**74.5**	**0.924**	**0.463**	$\mathbf{1.2 \times 10^4}$	**63.1**	**0.630**	0.586	$\mathbf{1.7 \times 10^4}$	**63.6**	**0.634**	**0.578**

Conclusion

- We propose a novel attribution method called FANS that can better distinguish the contribution of feature subsets to predictions.
- FANS defines a novel attribution as the highest probability in the Probabilities of being a Necessity and Sufficiency (PNS) cause of the prediction change for perturbing samples in different neighborhoods,
- We develop a dual-stage (factual and interventional) perturbation test to implement our method.