Explaining Cross-Modal Model Behavior with Gradient-Estimation-Based Feature Interaction

Yi Cai*†, *Nanyang Technological University*
Xinpeng Li*, *Nanjing University*
Gerhard Wunder, *Freie Universität Berlin*

Correspondence: **yi.cai[AT]ntu.edu.sg**

NEURAL INFORMATION PROCESSING SYSTEMS

* These authors contributed equally.
† Yi Cai was affiliated with Freie Universität Berlin when the work was carried out.

From Feature Attribution to Feature Interaction

Are individual contributions sufficient for explanations?

Feature Attribution

- First-order explanation

$$A_f(\boldsymbol{x}, \dot{\boldsymbol{x}}) \hookrightarrow \{\xi_i\}_{i=1}^{d} \in \mathbb{R}^d$$



From Feature Attribution to Feature Interaction

Are individual contributions sufficient for explanations?

Feature Attribution

- First-order explanation

$$A_f(\boldsymbol{x}, \dot{\boldsymbol{x}}) \hookrightarrow \{\xi_i\}_{i=1}^{d} \in \mathbb{R}^d$$



a white horse pulling a carriage with a man on it .



a white horse pulling a carriage with a man on it .

From Feature Attribution to Feature Interaction

Are individual contributions sufficient for explanations?

Feature Attribution

- First-order explanation

$$A_f(\boldsymbol{x}, \dot{\boldsymbol{x}}) \hookrightarrow \{\xi_i\}_{i=1}^{d} \in \mathbb{R}^d$$

Feature Interaction

- Higher-order explanation

$$I_f(\boldsymbol{x}, \dot{\boldsymbol{x}}, \boldsymbol{S}) \hookrightarrow \left\{\xi_{\boldsymbol{S},i}\right\}_{i \notin \boldsymbol{S}} \in \mathbb{R}^{d-|\boldsymbol{S}|}$$

a white horse pulling a carriage with a man on it .



a white horse pulling a carriage with a man on it .

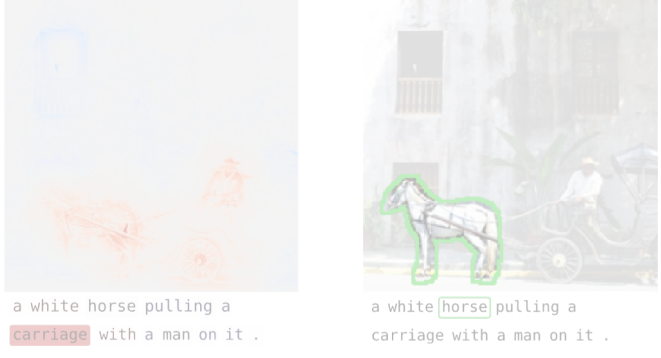
From Feature Attribution to Feature Interaction

Are individual contributions sufficient for explanations?

Feature Attribution

- First-order explanation

$$A_f(\boldsymbol{x}, \dot{\boldsymbol{x}}) \hookrightarrow \{\xi_i\}_{i=1}^{d} \in \mathbb{R}^d$$



Feature Interaction

- Higher-order explanation

$$I_f(\boldsymbol{x}, \dot{\boldsymbol{x}}, \boldsymbol{S}) \hookrightarrow \left\{\xi_{\boldsymbol{S},i}\right\}_{i \notin \boldsymbol{S}} \in \mathbb{R}^{d-|\boldsymbol{S}|}$$



Feature Interaction provides an ideal tool to examine cross-modal concept alignments.

From Attribution to Interaction

Interaction as a high-order extension of attribution

Define Proxy Gradient Estimator[1]

- Score-function estimator

$$\nabla J := \mathbb{E}_{z \sim \pi(\cdot|x)} \left[f(z) \cdot \nabla \log \pi(z|x) \right]$$

Search distribution

[1] Cai, Yi, Thibaud Ardoin, and Gerhard Wunder. "GEFA: A general feature attribution framework using proxy gradient estimation." Forty-second International Conference on Machine Learning. 2025.

From Attribution to Interaction

Interaction as a high-order extension of attribution

Define Proxy Gradient Estimator[1]

- Score-function estimator

$$\nabla J := \mathbb{E}_{z \sim \pi(\cdot|x)} \left[f(z) \cdot \nabla \log \pi(z|x) \right]$$

Search distribution

- Define π through proxy variables α

$$z = x \odot \epsilon + \dot{x} \odot \bar{\epsilon}, \qquad \epsilon \sim \text{Bernoulli}(\alpha)$$

[1] Cai, Yi, Thibaud Ardoin, and Gerhard Wunder. "GEFA: A general feature attribution framework using proxy gradient estimation." Forty-second International Conference on Machine Learning. 2025.

From Attribution to Interaction

Interaction as a high-order extension of attribution

Define Proxy Gradient Estimator[1]

- Score-function estimator

$$\nabla J := \mathbb{E}_{z \sim \pi(\cdot|x)} \left[f(z) \cdot \nabla \log \pi(z|x) \right]$$

Search distribution

- Define π through proxy variables α

$$z = x \odot \epsilon + \dot{x} \odot \bar{\epsilon}, \qquad \epsilon \sim \text{Bernoulli}(\alpha)$$

- Substitute π and estimate w.r.t. α:

$$\eta_i(x(\alpha)) = \mathbb{E}_{\pi(\epsilon|\alpha)} \left[f(z) \cdot \left(\frac{\epsilon_i - \alpha_i}{\alpha_i \, (1 - \alpha_i)} \right) \right]$$

[1] Cai, Yi, Thibaud Ardoin, and Gerhard Wunder. "GEFA: A general feature attribution framework using proxy gradient estimation." Forty-second International Conference on Machine Learning. 2025.

From Attribution to Interaction

Interaction as a high-order extension of attribution

Define Proxy Gradient Estimator[1]

- Score-function estimator

$$\nabla J := \mathbb{E}_{z\sim\pi(\cdot|x)}\left[f(z)\cdot\nabla\log\pi(z|x)\right]$$

Search distribution

- Define π through proxy variables α

$$z = x\odot\epsilon + \dot{x}\odot\bar{\epsilon}, \qquad \epsilon\sim\text{Bernoulli}(\alpha)$$

- Substitute π and estimate w.r.t. α:

$$\eta_i(x(\alpha)) = \mathbb{E}_{\pi(\epsilon|\alpha)}\left[f(z)\cdot\left(\frac{\epsilon_i-\alpha_i}{\alpha_i\,(1-\alpha_i)}\right)\right]$$

Extension to GEFI for Feature Interaction

- k-th-order Proxy Gradient Estimator

$$\eta_{S,i}(x(\alpha)) = \mathbb{E}_{\pi(\epsilon|\alpha)}\left[f(z)\cdot\left(\prod_{k\in S\cup\{i\}}\frac{\epsilon_k-\alpha_k}{\alpha_k\,(1-\alpha_k)}\right)\right]$$

Mixed partial derivatives

From Attribution to Interaction

Interaction as a high-order extension of attribution

Define Proxy Gradient Estimator[1]

- Score-function estimator

$$\nabla J := \mathbb{E}_{z\sim\pi(\cdot|x)}\left[f(z)\cdot\nabla\log\pi(z|x)\right]$$

Search distribution

- Define π through proxy variables α

$$z = x\odot\epsilon + \dot{x}\odot\bar{\epsilon}, \qquad \epsilon\sim\text{Bernoulli}(\alpha)$$

- Substitute π and estimate w.r.t. α:

$$\eta_i(x(\alpha)) = \mathbb{E}_{\pi(\epsilon|\alpha)}\left[f(z)\cdot\left(\frac{\epsilon_i-\alpha_i}{\alpha_i\,(1-\alpha_i)}\right)\right]$$

Extension to GEFI for Feature Interaction

- k-th-order Proxy Gradient Estimator

$$\eta_{S,i}(x(\alpha)) = \mathbb{E}_{\pi(\epsilon|\alpha)}\left[f(z)\cdot\left(\prod_{k\in S\cup\{i\}}\frac{\epsilon_k-\alpha_k}{\alpha_k\,(1-\alpha_k)}\right)\right]$$

Mixed partial derivatives

- Integrate over a straightline path $\alpha = \gamma\cdot\mathbf{1}_d$

$$\xi_S = \int_0^1 \eta_S(x(\gamma\cdot\mathbf{1}_d))\,\mathrm{d}\gamma$$

From Attribution to Interaction

Interaction as a high-order extension of attribution

Define Proxy Gradient Estimator[1]

- Score-function estimator

$$\nabla J := \mathbb{E}_{z\sim\pi(\cdot|x)}\left[f(z)\cdot\nabla\log\underbrace{\pi(z|x)}_{\text{Search distribution}}\right]$$

- Define π through proxy variables α

$$z = x\odot\epsilon + \dot{x}\odot\bar{\epsilon},\qquad \epsilon\sim\text{Bernoulli}(\alpha)$$

- Substitute π and estimate w.r.t. α:

$$\eta_i(x(\alpha)) = \mathbb{E}_{\pi(\epsilon|\alpha)}\left[f(z)\cdot\left(\frac{\epsilon_i-\alpha_i}{\alpha_i\,(1-\alpha_i)}\right)\right]$$

Extension to GEFI for Feature Interaction

- k-th-order Proxy Gradient Estimator

$$\eta_{S,i}(x(\alpha)) = \mathbb{E}_{\pi(\epsilon|\alpha)}\left[f(z)\cdot\left(\underbrace{\prod_{k\in S\cup\{i\}}\frac{\epsilon_k-\alpha_k}{\alpha_k\,(1-\alpha_k)}}_{\text{Mixed partial derivatives}}\right)\right]$$

- Integrate over a straightline path $\alpha = \gamma\cdot\mathbf{1}_d$

$$\xi_S = \int_0^1 \eta_S(x(\gamma\cdot\mathbf{1}_d))\,\mathrm{d}\gamma$$

GEFI is mathematically equivalent to SII[2]

$$\xi_S = \int_0^1 \eta_S\big(x(\gamma\cdot\mathbf{1}_d)\big)\,\mathrm{d}\gamma \Leftrightarrow \sum_{T\subseteq[d]\setminus S}\frac{|T|!\,(d-|T|-k)!}{(d-k+1)!}\cdot\delta_S(T)$$

[2] Collapse to SV when $|S| = 0$.

Patch-wise vs. Pixel-wise Interactions

The formulation in GEFI brings computational advantage by avoiding the difference operator $\delta_S(\boldsymbol{T})$

- Simultaneous estimation for all features
- Reusable observations for different subsets.

Patch-wise vs. Pixel-wise Interactions

The formulation in GEFI brings computational advantage by avoiding the difference operator $\delta_S(\boldsymbol{T})$

- Simultaneous estimation for all features
- Reusable observations for different subsets.



Finer granularity of interaction results: Pixel-wise interaction

Low computational costs compared to SOTA

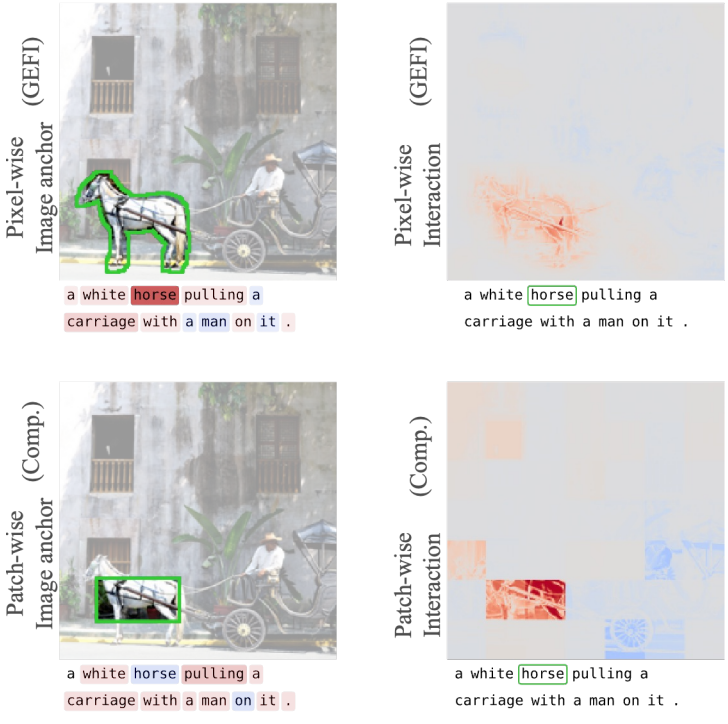
Patch-wise vs. Pixel-wise Interactions

The formulation in GEFI brings computational advantage by avoiding the difference operator $\delta_S(T)$

- Simultaneous estimation for all features
- Reusable observations for different subsets.

Finer granularity of interaction results: Pixel-wise interaction

Low computational costs compared to SOTA



THANK YOU



Code & Resources

Read the paper

Correspondence: **yi.cai[AT]ntu.edu.sg**