How to unify attribution explanations by interactions?

Huiqi Deng

Attribution definition

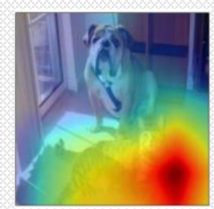
Attribution explanation

- A branch of semantic explanations
- Inferring contribution score of each individual feature

Definition 1. *For a pre-trained model f, an attribution of prediction at input $\boldsymbol{x} = [x_1, \ldots, x_n]$ is a vector $\boldsymbol{a} = [a_1, \ldots, a_n]$, where a_i is the contribution of x_i to the prediction $f(\boldsymbol{x})$.*

Input image

Predict ⇒ Label: cat ⇐ *Interpret*



Attribution Heatmap

[1] Sundararajan et al. "Axiomatic attribution for deep networks." ICML, 2017.

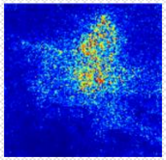
Existing attribution methods

Many attribution methods are proposed recently.

- **Sensitivity**
- **Perturbation**
- **Layerwise decomposition**
- **Averaging gradients**

Various heuristics

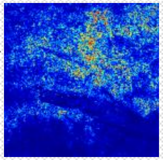
Expected Gradients



$\int_{\tilde{\boldsymbol{x}}} p_D(\tilde{\boldsymbol{x}}) a^{IG} d\tilde{\boldsymbol{x}}$

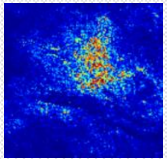
......

*Gradient *Input*



$f_{x_i}(\boldsymbol{x}) x_i$

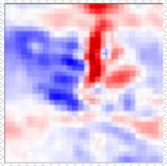
Integrated Grads



$(x_i - \tilde{x}_i) \int_{\boldsymbol{p}} f_{x_i}(\boldsymbol{x}) d\boldsymbol{p}$

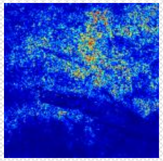
Input sample

Occlusion



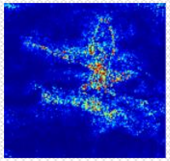
$f(\boldsymbol{x}) - f(\boldsymbol{x}|_{\boldsymbol{p_j}=\boldsymbol{0}})$

ϵ-LRP



$\frac{z_{ji}^{(l)}}{\sum_{i'} z_{ji'}^{(l)}} f(\boldsymbol{z})$

DeepLIFT



$\frac{z_{ji}^{(l)} - \tilde{z}_{ji}^{(l)}}{\sum_{i'} z_{ji'}^{(l)} - \sum_{i'} \tilde{z}_{ji'}^{(l)}} \Delta f$

Different formulations

Problems of attribution explanation

- The attrbution problem is not well-defined
  - The definition is uninformative for how to assign the contribution
- Many attribution methods are based on different heurisitcs
  - Few theoretical foundations
  - No mutuality among existing methods
  - Difficult to compare theoretically

IJCAI-21 TUTORIALS
#IJCAI2021

Contributions of this paper

- We propose a **Taylor attribution framework**, which offers a theoretical formulation to the attribution problem.

- *Fourteen* mainstream attribution methods with different formulations are **unified into** the proposed framework by theoretical reformulations.

- We propose principles for ***a reasonable attribution***, and **assess the fairness** of existing attribution methods.

Deng et al. A Unified Taylor Framework for Revisiting Attribution Methods, AAAI, 2021.
Deng et al. A General Taylor Framework for Unifying and Revisiting Attribution Methods. in arXiv:2105.13841

Contributions of this paper

- We propose a **Taylor attribution framework**, which offers a theoretical formulation for how to assign contribution.
- Fourteen mainstream attribution methods are unified into the proposed Taylor framework by theoretical reformulations.
- We propose principles for **a reasonable attribution**, and **assess the fairness** of existing attribution methods.

Deng et al. A Unified Taylor Framework for Revisiting Attribution Methods, AAAI, 2021.
Deng et al. A General Taylor Framework for Unifying and Revisiting Attribution Methods. in arXiv:2105.13841

Attribution problem statement

Input: pre-trained model f, input sample $\boldsymbol{x}$, and baseline $\tilde{\boldsymbol{x}}$ (*no signal state*)

Output: attribution vector $\boldsymbol{a}$

Many attribution methods aim to **distribute the outcome of x (w.r.t the baseline $\tilde{x}$) to each feature**,

$$f(\boldsymbol{x}) - f(\tilde{\boldsymbol{x}}) = a_1 + \ldots + a_n$$

$\downarrow$

Corresponds to $v(N) - v(\emptyset)$

However, there are infinite possible cases for such decomposition.

Which decomposition is reasonable?

Deng et al. A Unified Taylor Framework for Revisiting Attribution Methods, AAAI, 2021.
Deng et al. A General Taylor Framework for Unifying and Revisiting Attribution Methods. in arXiv:2105.13841

Taylor attribution framework

☐ Challenges

- DNN Model f is too complex to analyze

☐ Basic idea

- Taylor Theroem: If $f(\boldsymbol{x})$ is infinitely differentiable, then $f(\boldsymbol{x}) - f(\widetilde{\boldsymbol{x}})$ can be approximated by a Taylor expansion function
- The Taylor expansion function can be explictly divided into independent and interactive parts
- Then the attribution can be expressed as a function of Taylor independent and interaction terms

Deng et al. A Unified Taylor Framework for Revisiting Attribution Methods, AAAI, 2021.
Deng et al. A General Taylor Framework for Unifying and Revisiting Attribution Methods. in arXiv:2105.13841

Second-order Taylor attribution

- Second-order Taylor expansion

$$f(\boldsymbol{x}) - f(\tilde{\boldsymbol{x}}) = \sum_i f_{x_i} \Delta_i + \frac{1}{2}\sum_i \sum_j f_{x_i x_j} \Delta_i \Delta_j + \boldsymbol{\varepsilon} \qquad \mathbf{(1)}$$

- Divide the expansion into first-order, high-order independent and interaction terms

$$f(\boldsymbol{x}) - f(\tilde{\boldsymbol{x}}) = \underbrace{\sum_i f_{x_i} \Delta_i}_{\text{All first-order terms, } \boldsymbol{T_i^{\alpha}}} + \underbrace{\frac{1}{2}\sum_i f_{x_i{}^2} \Delta_i{}^2}_{\text{All high-order independent terms, } \boldsymbol{T_i^{\gamma}}} + \underbrace{\frac{1}{2}\sum_{i \neq j} f_{x_i x_j} \Delta_i \Delta_j}_{\text{All high-order interaction terms } \boldsymbol{I}(\boldsymbol{S})} + \boldsymbol{\varepsilon} \qquad \mathbf{(2)}$$

- Attribution vector can be expressed as a function of the three type terms

$$a_i = decompose(f(\boldsymbol{x}) - f(\tilde{\boldsymbol{x}})) \Longrightarrow a_i = \varphi(T_i^{\alpha}, T_i^{\gamma}, I(S))$$

Connections with related work in Game theory

- Connection to Shapley Taylor interaction index [1]
  - Shapely Taylor interaction index $J^k(S)$ measures Taylor interactions of subsets with at most k players.
  - **When $k = n$**, i.e., consider interactions of all subsets,

$$J^n(S) = I(S), \quad \forall S$$

  - $I(S)$ is **a special case** of Shapley Taylor interaction index.

[1] Sundararajan, et al. The shapley taylor interaction index. ICML, 2020.

Contributions of this paper

- We propose a **Taylor attribution framework**, which offers a theoretical formulation for the attribution problem.
- We prove that, *Fourteen* attribution methods with different formulas can be **unified into** the proposed Taylor attribution framework.
- We propose principles for **a reasonable attribution**, and **assess the fairness** of existing attribution methods.

Deng et al. A Unified Taylor Framework for Revisiting Attribution Methods, AAAI, 2021.
Deng et al. A General Taylor Framework for Unifying and Revisiting Attribution Methods. in arXiv:2105.13841

Unifying attribution maps of fourteen methods by interactions

Attribution maps of *Fourteen* methods are unified into the Taylor attribution framework. Specifically, they can expressed as a weighted sum of the three type terms.

$$\boldsymbol{a_i = \alpha_i T_i^{\alpha} + \gamma_i T_i^{\gamma} + \sum_S c_i^S I(S)}$$

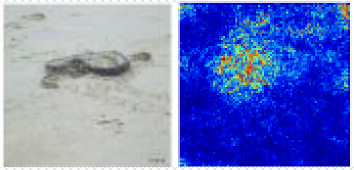
$\alpha_i, \gamma_i, c_i^S$ **are the coefficients**

Categorization	Methods	Taylor Reformulations
Basic versions	GI [21]	$a_i^{GI} = T_i^{\alpha}$
	LRP-ϵ [8]	$a_i^{LRP\epsilon} = T_i^{\alpha}$
	GCAM [31]	$a_{ij}^{GCAM} = (T^{\alpha}(h))_{ij}$
	Occ-1 [19]	$a_i^{Occ1} = T_i^{\alpha} + T_i^{\gamma_d} + \sum_{\{i\} \subsetneq A} T_A^{\gamma_t}$
	Occ-p [22]	$a_i^{Occp} = T_{\boldsymbol{p_j}}^{\alpha} + T_{\boldsymbol{p_j}}^{\gamma_d} + \sum_{\boldsymbol{p_j} \cap A \neq \emptyset} T_A^{\gamma_t}$
	Integrated [7]	$a_i^{IG} = T_i^{\alpha} + T_i^{\gamma_d} + a_i^{\gamma_t}(IG)$
	DeepLIFT [23]	$a_i^{DL}(l) = a_i^{IG}(l) = T_i^{\alpha} + T_i^{\gamma_d} + a_i^{\gamma_t}(IG)$
	Shapley [29]	$a_i^{Shap} = T_i^{\alpha} + T_i^{\gamma_d} + a_i^{\gamma_t}(Shap)$
Separating + & -	DeepLIFT+- [23]	$a_i^{DL+} = T_i^{\alpha} + T_i^{\gamma_d} + a_i^{\gamma_t}(DL+)$ $a_i^{DL-} = T_i^{\alpha} + T_i^{\gamma_d} + a_i^{\gamma_t}(DL-)$
	Deep Taylor [25]	$a_i^{DTD} = T_i^{\alpha} + T_i^{\gamma_d} + a_i^{\gamma_t}(IG) + cT_{N-}$
	LRP-$\alpha\beta$ [8]	$a_i^{+} = \alpha(T_i^{\alpha} + T_i^{\gamma_d} + a_i^{\gamma_t}(IG) + cT_{N-})$ $a_i^{-} = -\beta(T_i^{\alpha} + T_i^{\gamma_d} + a_i^{\gamma_t}(IG) + cT_{N+})$
Expected Attribution	Expected Grads [28]	$a_i^{EG} = \int_{\tilde{\boldsymbol{x}}} p_D(\tilde{\boldsymbol{x}}) a_i^{IG} d\tilde{\boldsymbol{x}}$
	Expected DeepLIFT	$a_i^{EDL} = \int_{\tilde{\boldsymbol{x}}} p_D(\tilde{\boldsymbol{x}}) a_i^{DL} d\tilde{\boldsymbol{x}}$
	Deep Shapley [5]	$a_i^{DShap} \approx \int_{\tilde{\boldsymbol{x}}} p_D(\tilde{\boldsymbol{x}}) a_i^{Shap} d\tilde{\boldsymbol{x}}$

Unifying attribution maps of fourteen methods by interactions: Gradient×Input

Intuition. Gradient*Input produces attribution maps with improved sharpness, by multipling the gradients with the input.

Unification. *Gradient×Input can be unified into Taylor attribution framework.*



Reformulation. In Gradient×Input, the corresponding coefficients are,

$$\alpha_i = 1, \qquad \text{First-order terms, } \boldsymbol{T_i^\alpha}$$

$$\gamma_i = 0, \qquad \text{high-order independent terms, } \boldsymbol{T_i^\gamma}$$

$$c_i^S = 0, \quad \forall S \qquad \text{high-order interaction terms, } \boldsymbol{I(S)}$$

- Only assigns the first-order terms

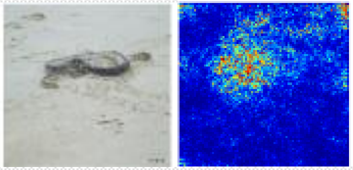
$$\boldsymbol{a_i = \alpha_i T_i^\alpha + \gamma_i T_i^\gamma + \sum_S c_i^S I(S)}$$

Deng et al. A Unified Taylor Framework for Revisiting Attribution Methods, AAAI, 2021.
Deng et al. A General Taylor Framework for Unifying and Revisiting Attribution Methods. in arXiv:2105.13841

Unifying attribution maps of fourteen methods by interactions: ε-LRP

Intuition. It produces *attribution maps by* distributing the output in proportion according to the input. It conducts in a layer-wise manner.

Unification. *ε-LRP can be unified into the Taylor attribution framework.*



Reformulation. In ε-LRP, if relu is used as activation function, the corresponding coefficients are [1],

$\alpha_i = 1,$	First-order terms, $\boldsymbol{T_i^{\alpha}}$
$\gamma_i = 0,$	high-order independent terms, $\boldsymbol{T_i^{\gamma}}$
$c_i^S = 0, \quad \forall S$	high-order interaction terms, $\boldsymbol{I}(\boldsymbol{S})$

- Only assigns the first-order terms when relu is applied.

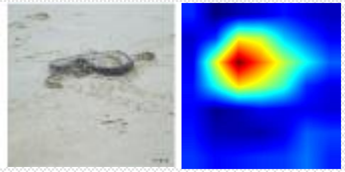
$$a_i = \alpha_i T_i^{\alpha} + \gamma_i T_i^{\gamma} + \sum_S c_i^S I(S)$$

[1] Ancona, Marco, et al. Towards better understanding of gradient-based attribution methods for deep neural networks. ICLR, 2018.

Unifying attribution maps of fourteen methods by interactions: GradCAM

Intuition. GradCAM conducts global average pooling to the gradients, then perform a linear combination

Unification. *GradCAM can be unified into Taylor attribution framework.*



Reformulation. Define the global average pooled features as F. Consider $f(\boldsymbol{x}) = h(F)$. Then in GradCAM, the corresponding coefficients **of function $\boldsymbol{h}$** are,

$$\alpha_i = 1, \quad \text{First-order terms, } \boldsymbol{T_i^\alpha}$$

$$\gamma_i = 0, \quad \text{high-order independent terms, } \boldsymbol{T_i^\gamma}$$

$$c_i^S = 0, \quad \forall S \quad \text{high-order interaction terms, } \boldsymbol{I}(\boldsymbol{S})$$

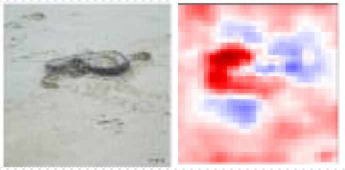
- Assigns the first-order terms of function h.

$$\boldsymbol{a_i} = \boldsymbol{\alpha_i T_i^\alpha} + \boldsymbol{\gamma_i T_i^\gamma} + \sum_{\boldsymbol{S}} \boldsymbol{c_i^S I(S)}$$

Unifying attribution maps of fourteen methods by interactions: Occlusion-1&patch

Intuition. Occlude one pixel/patch, and observe how the prediction changes.

Unification. *Occlusion-1 & Occlusion-patch can be unified into Taylor framework.*



Reformulation. In Occlusion-1, the corresponding coefficients are,

$$\alpha_i = 1, \qquad \text{First-order terms, } \boldsymbol{T_i^{\alpha}}$$

$$\gamma_i = 1, \qquad \text{high-order independent terms, } \boldsymbol{T_i^{\gamma}}$$

$$\left.\begin{aligned} c_i^S &= 1, \quad if\ i \in S \\ c_i^S &= 0, \quad if\ i \notin S \end{aligned}\right\} \quad \text{high-order interaction terms, } \boldsymbol{I(S)}$$

- assigns first-order, high-order independent terms of x_i, and all interactions involving x_i.

$$\boldsymbol{a_i = \alpha_i T_i^{\alpha} + \gamma_i T_i^{\gamma} + \sum_S c_i^S I(S)}$$

Unifying attribution maps of fourteen methods by interactions: Shapley value

Intuition. Shapley value obtains the attribution map by averaging the marginal contribution of x_i to coalition S over all possible coalitions involving x_i.

Unification. *Shapley value can be unified into Taylor attribution framework.*

Reformulation. In Shapley value, the corresponding coefficients are,

$\alpha_i = 1,$ First-order terms, $\boldsymbol{T_i^\alpha}$

$\gamma_i = 1,$ high-order independent terms, $\boldsymbol{T_i^\gamma}$

$$\left.\begin{array}{ll} c_i^S = 1/|S|, & if\ i \in S \\ c_i^S = 0, & if\ i \notin S \end{array}\right\} \text{high-order interaction terms, } \boldsymbol{I(S)}$$

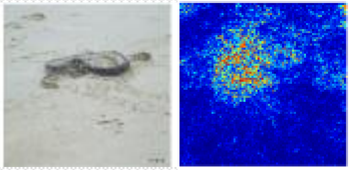
- assigns first-order, independent terms of x_i, and $1/|S|$ proportion of interactions involving x_i.

$$\boldsymbol{a_i} = \boldsymbol{\alpha_i T_i^\alpha} + \boldsymbol{\gamma_i T_i^\gamma} + \sum_{S} \boldsymbol{c_i^S I(S)}$$

Unifying attribution maps of fourteen methods by interactions: Integrated Grads

Intuition. It produces the attribution map by integrating the gradients along a straight line from baseline $\tilde{\boldsymbol{x}}$ to input $\boldsymbol{x}$.

Unification. *Integrated Gradients can be unified into Taylor attribution framework.*



Reformulation. In Integrated Gradients, the corresponding coefficients are,

$$\alpha_i = 1, \qquad \text{First-order terms, } \boldsymbol{T_i^{\alpha}}$$

$$\gamma_i = 1, \qquad \text{high-order independent terms, } \boldsymbol{T_i^{\gamma}}$$

$$\left.\begin{array}{ll} c_i^S(\pi) = k_i/K, & if\ i \in S, \pi = [k_1, \ldots, k_n], \\ c_i^S = 0, & if\ i \notin S \end{array}\right\} \text{high-order interaction terms, } \boldsymbol{I(S)}$$

$$K = k_1 + \ldots + k_n$$

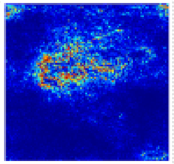
- assigns first-order,independent terms of x_i, and $\boldsymbol{k_i/K}$ proportion of interaction terms $\boldsymbol{x_1^{k_1} x_2^{k_2} \ldots x_n^{k_n}}$ to x_i.
- For example, $f(\boldsymbol{x}) = x_1 x_2^3 + x_1^2 x_2 x_3^2$, then $a_2 = \frac{3}{4} x_1 x_2^3 + \frac{1}{5} x_1^2 x_2 x_3^2$.

$$a_i = \alpha_i T_i^{\alpha} + \gamma_i T_i^{\gamma} + \sum_S c_i^S I(S)$$

Unifying attribution maps of fourteen methods by interactions: DeepLIFT Rescale

Intuition. DeepLIFT propogates the output difference in proportion according to the input difference. Such propogation proceeds in a layer-wise manner.

Unification. *DeepLIFT Rescale can be unified into Taylor attribution framework.*



Reformulation. Consider the attribution at l layer. If $\boldsymbol{f_l(z)} = \boldsymbol{\sigma(w^T z + b)}$, then in DeepLIFT Rescale, the corresponding coefficients are,

$$\alpha_i = 1,$$ First-order terms, $\boldsymbol{T_i^\alpha}$

$$\gamma_i = 1,$$ high-order independent terms, $\boldsymbol{T_i^\gamma}$

$$\left.\begin{array}{ll} c_i^S(\pi) = k_i/K, & if\ i \in S, \pi = [k_1, \ldots, k_n] \\ c_i^S = 0, & if\ i \notin S \end{array}\right\}$$

$K = k_1 + \ldots + k_n$

high-order interaction terms, $\boldsymbol{I(S)}$

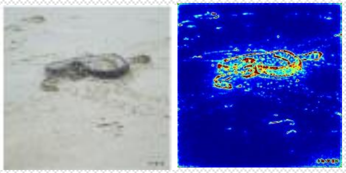
- Shares the same coefficients as Integrated gradients at each layer.

$$\boldsymbol{a_i = \alpha_i T_i^\alpha + \gamma_i T_i^\gamma + \sum_S c_i^S I(S)}$$

Unifying attribution maps of fourteen methods by interactions: Deep Taylor

Intuition. proceeds in a layer-wise manner. It propgates all relevances to the features with positive weight.

Unification. *Deep Taylor can be unified into the framework.*



Reformulation. Define $N^+ = \{i|w_{ji} \geq 0\}$ and $N^- = \{i|w_{ji} < 0\}$, where w_{ji} is the parameters at l layer. In Deep Taylor, for features in N^+, the coefs are,

$$\alpha_i = 1, \qquad \text{First-order terms, } \boldsymbol{T_i^\alpha}$$

$$\gamma_i = 1, \qquad \text{high-order independent terms, } \boldsymbol{T_i^\gamma}$$

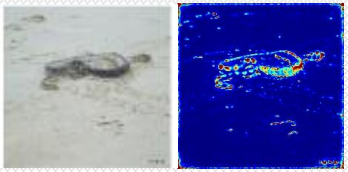
$$\left.\begin{aligned} c_i^S(\pi) &= k_i/K, && if\ i \in S, \pi = [k_1, \ldots, k_n] \\ c_i^S &= 0, && if\ i \notin S, S \subset N^- \\ \boxed{c_i^S = z_{ji}{}^+/z_j,} && && \boxed{if\ i \notin S, S \subseteq N}^- \end{aligned}\right\} \quad K = k_1 + \ldots + k_n \quad \text{high-order interaction terms, } \boldsymbol{I(S)}$$

- Noted that the interactions among features in N^- are assigned to features in N^+.

$$a_i = \alpha_i T_i^\alpha + \gamma_i T_i^\gamma + \sum c_i^S I(S)$$

Unifying attribution maps of fourteen methods by interactions: LRP-$\alpha\beta$

Intuition. propgates $\boldsymbol{\alpha}$ times relevances to the features with positive weight, and $\boldsymbol{\beta}$ times to the features with negative weight.



Unification. *LRP-$\alpha\beta$ can be unified into the Taylor attribution framework.*

Reformulation. Define $N^+ = \{i|w_{ji} \geq 0\}$ and $N^- = \{i|w_{ji} < 0\}$, where w_{ji} is the parameters at l layer. In LRP-$\alpha\beta$, for features in N^+, the coefficients are,

$$\alpha_i = \alpha,$$

First-order terms, $\boldsymbol{T_i^\alpha}$

$$\gamma_i = \alpha,$$

high-order independent terms, $\boldsymbol{T_i^\gamma}$

$$c_i^S(\pi) = \alpha\, k_i/K\,,$$

$$if\ i \in S, \pi = [k_1, \ldots, k_n]$$

$$c_i^S = 0, \quad if\ i \notin S, \quad S \subset N^-$$

$$K = k_1 + \ldots + k_n$$

high-order interaction terms, $\boldsymbol{I(S)}$

$$c_i^S = \alpha\, z_{ji}^+/z_j\ if\ i \notin S, S \subseteq N^-$$

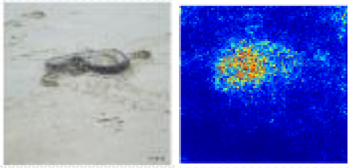
- The coefficients are α times the coefficients in Deep Taylor attribution.

$$a_i = \alpha_i T_i^\alpha + \gamma_i T_i^\gamma + \sum_S c_i^S I(S)$$

Unifying attribution maps of fourteen methods by interactions: Expected Attribution

Intuition. proposed to reduce the probability that attribution is dominated by a specific baseline, which averages the attributions over multiple baselines.

$$a_i^{exp} = \int p(\tilde{\boldsymbol{x}}) a_i^{basic} d\tilde{\boldsymbol{x}} \qquad (1)$$



where a_i^{basic} is the attribution obtained by basic methods.

Unification. Combining Eq.(1) with the previous reformulations, Expected Attributions can be unified into the Taylor attribution framework.

- For example, Expected Gradients, Expected DeepLIFT, and Deep Shapley.

$$\boldsymbol{a_i = \alpha_i T_i^{\alpha} + \gamma_i T_i^{\gamma} + \sum_S c_i^S I(S)}$$

Contributions of this paper

- We propose a **Taylor attribution framework**, which offers a theoretical formulation for the attribution problem.
- We prove that, *Fourteen* attribution methods with different formula can be **unified into** the proposed Taylor attribution framework.
- We propose principles for ***a reasonable attribution***, and **assess the fairness** of existing attribution methods.

Deng et al. A Unified Taylor Framework for Revisiting Attribution Methods, AAAI, 2021.
Deng et al. A General Taylor Framework for Unifying and Revisiting Attribution Methods. in arXiv:2105.13841

Principles for a reasonable attribution

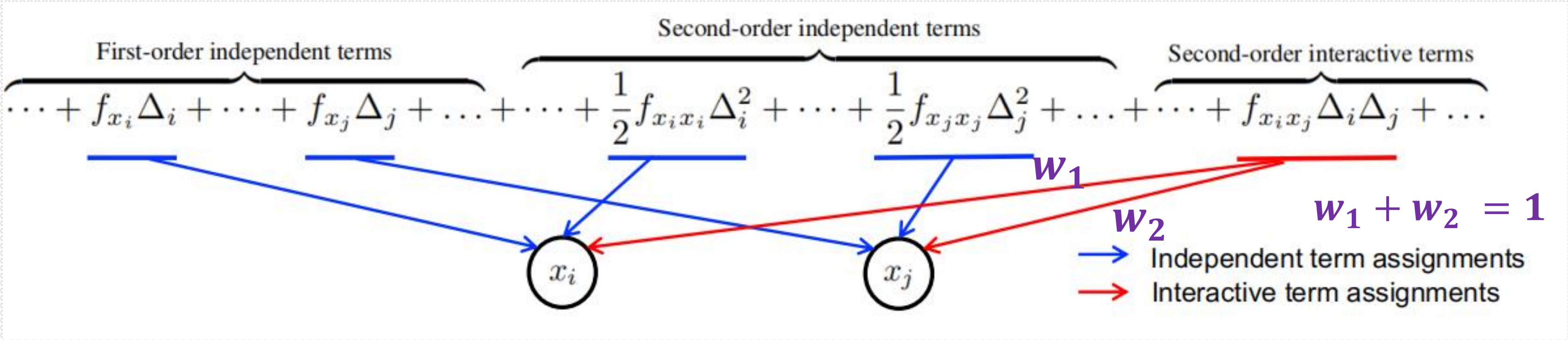
- We proved that, attribution maps of fourteen methods can be unified as the following form:

$$a_i = \alpha_i T_i^{\alpha} + \gamma_i T_i^{\gamma} + \sum_{S} c_i^S I(S)$$

How to define a reasonable attribution map?

Deng et al. A Unified Taylor Framework for Revisiting Attribution Methods, AAAI, 2021.
Deng et al. A General Taylor Framework for Unifying and Revisiting Attribution Methods. in arXiv:2105.13841

Principles for a reasonable attribution



Principle 1:

- The first-order terms of x_i should be all assigned to x_i.
- The high-order independent terms of x_i should be all assigned to x_i.
- Only Interactions of S involving x_i , should be assigned to x_i.

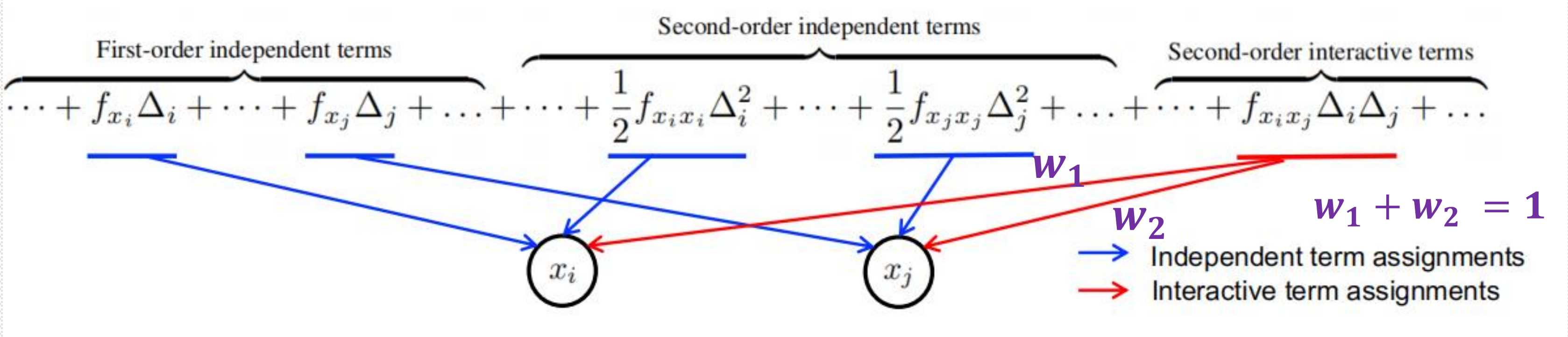
$$\alpha_i = 1, \gamma_i = 1,$$

$$c_i^S > 0, \quad if\ i \in S$$

$$c_i^S = 0, \quad if\ i \notin S$$

$$a_i = \alpha_i T_i^{\alpha} + \gamma_i T_i^{\gamma} + \sum_S c_i^S I(S)$$

Principles for a reasonable attribution



Principle 2:

- Interactions of any coalition S should be all distributed to the players in S.

$$\sum_{i \in S} c_i^S = 1, \qquad \forall S$$

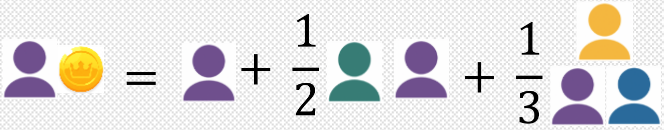
$$a_i = \alpha_i T_i^{\alpha} + \gamma_i T_i^{\gamma} + \sum_{S} c_i^S I(S)$$

Assessing the fairness of existing attribution methods

These principles can be applied to assess the fairness of exsiting methods.

➢ For example, Shapley value well satisfies the two principles.

$$\alpha_i = 1, \gamma_i = 1,$$
$$c_i^S = 1/|S|, \quad if\ i \in S$$
$$c_i^S = 0, \quad if\ i \notin S$$



- Interactions of S are evenly assigned to the players in S.
- In this sense, Shapley value is a **fair** attribution.

➢ For example, Occlusion-1 satisfies principle 1, doesn't satisfy principle 2.

$$\alpha_i = 1, \gamma_i = 1,$$
$$c_i^S = 1, \quad if\ i \in S$$
$$c_i^S = 0, \quad if\ i \notin S$$
$$\Longrightarrow \sum_{i \in S} c_i^S = |S|, \quad \forall S$$

- Interactions of S are repeatedly assigned to each player.