Towards Attributions of Input Variables in a Coalition

International Conference on Machine Learning (ICML 2025)

Xinhao Zheng, Huiqi Deng, Quanshi Zhang

Shanghai Jiao Tong University

Attribution methods

Conflict of attributions

Definition 3.1. *Given two partitions of n input variables $N = \{1, 2, ..., n\}$ and $P = \{S_1, S_2, ..., S_m\}$, subject to $N = \bigcup_{i=1}^{m} S_i$, $\forall i \neq j$, $S_i \cap S_j = \emptyset$, the conflict of attributions means that there exists a coalition S_k such that the attribution of the coalition S_k is not equal to the sum of attributions of its compositional variables,* i.e. $\phi_P(S_k) \neq \sum_{i \in S_k} \phi_N(i)$

Table 1. Comparison between the solutions of the conflict of attributions in different attribution methods

Attribution methods	Solutions for the conflict of attributions
Shapley value (Shapley et al., 1953)	Efficiency axiom $v(N) = \sum_{i \in N} \phi(i)$, but cannot ensure the efficiency property, *w.r.t.* any arbitrary set $S \subseteq N$, *i.e.*, $\varphi(S) \neq \sum_{i \in S} \phi(i)$
Banzhaf value (Penrose, 1946)	2-efficiency axiom: $B(i) + B(j) = B(\{i, j\})$ but do not satisfy $B(S) = \sum_{i \in S} B(i)$
Joint Shapley value (Harris et al., 2021)	Joint linearity, dummy, efficiency, anonymity, symmetry axioms, but estimating the attribution of a set of features/interactions, like (Sundararajan et al., 2020)
Faith-Shap (Tsai et al., 2023)	Using a loss $\|v(S) - \sum_{i \in S} \phi(i)\|^2$ to alleviate the conflict
Our method	**Proving the conflict is naturally unavoidable, and quantifying the essential cause for the conflict**

Attribution value for a coalition

Reformulating attributions

Theorem 3.2. ***(Reformulation of the Shapley value, proved in Appendix C)*** *The Shapley value $\phi(i)$ of each input variable x_i can be explained as $\phi(i) = \sum_{S\subseteq N, i\in S} \frac{1}{|S|}\left[I_{and}(S) + I_{or}(S)\right]$.*

Theorem 3.3. ***(Reformulation of the Banzhaf value, proved in Appendix D)*** *The Banzhaf value $B(i)$ of each input variable x_i can be reformulated as $B(i) = \sum_{S\subseteq N, i\in S} \frac{1}{2^{|S|-1}}\left[I_{and}(S) + I_{or}(S)\right]$.*

Attribution of a coalition

$$\forall S \subseteq N,\ \varphi(S) = \sum_{T\supseteq S} \frac{|S|}{|T|}\left[I_{\text{and}}(T) + I_{\text{or}}(T)\right]$$

Explaining the conflict of attributions

Theorem 3.4. *(proved in Appendix E) For any coalition $S \subseteq N$, we have $\sum_{i\in S}\phi(i) = \phi_{shared}(S) + \phi_{conflict}(S)$. $\phi_{shared}(S) \overset{def}{=} \varphi(S)$ is the attribution component existing in both the coalition's attribution $\varphi(S)$ and individual input variable's attribution $\phi(i)$, thereby being termed the* shared attribution component. *$\phi_{conflict}(S) = \sum_{T\subseteq N, T\cap S\neq\emptyset, T\cap S\neq S} \frac{|T\cap S|}{|T|}\left[I_{and}(T) + I_{or}(T)\right]$ represents the conflict (or difference) between the coalition attribution and the individual variables' attribution.*

The conflict of attributions comes from numerical effects of all interactions T that contain just partial but not all variables in S

Faithfulness of a coalition

Whether $U_{i,S}$ dominates the major effect of $\phi(i)$

$$R(i) = \frac{|U_{i,S}|}{|U_{i,S}| + |U_{i,\bar{S}}|}, \quad i \in S$$

Significance of the variable i participating in S

$$R'(i) = \frac{\sum_{T \supseteq S} \frac{1}{|T|}(|I_{\text{and}}(T)| + |I_{\text{or}}(T)|)}{\sum_{T' \ni i} \frac{1}{|T'|}(|I_{\text{and}}(T')| + |I_{\text{or}}(T')|)}, \quad i \in S$$

Significance of the entire coalition S

$$Q(S) = \frac{\sum_{T \supseteq S} \frac{|S|}{|T|}(|I_{\text{and}}(T)| + |I_{\text{or}}(T)|)}{\sum_{T' \subseteq N, T' \cap S \neq \emptyset} \frac{|T' \cap S|}{|T'|}(|I_{\text{and}}(T')| + |I_{\text{or}}(T')|)}$$

Experiments on toy functions

$$f(x) = \sum_{i=1}^{m} w_i \prod_{j \in T_i} x_j$$

where $x = [x_1, x_2, ..., x_n] \in \{0, 1\}^n, \forall i \neq j, T_i \neq T_j$.

For coalition S,

(1) purely faithful coalitions

$$\exists\, i, T_i \supseteq S \wedge \forall j (j \neq i), T_j \cap S = \emptyset$$

(2) partially faithful coalitions

$$\exists\, i, T_i \supseteq S \wedge (\exists\, i, T_i \cap S \neq \emptyset \wedge T_i \cap S \neq S)$$

(3) purely unfaithful coalitions $\quad$ *others*

Table 4. Coalition faithfulness metrics on toy functions

	$\mathbb{E}_{f,i}[R(i)]$	$\mathbb{E}_{f,i}[R'(i)]$	$\mathbb{E}_f[Q(S)]$
purely faithful coalitions	0.944	0.936	0.948
partially faithful coalitions	0.471	0.608	0.590
purely unfaithful coalitions	0.031	0.016	0.013

Experimental Results of faithfulness metrics

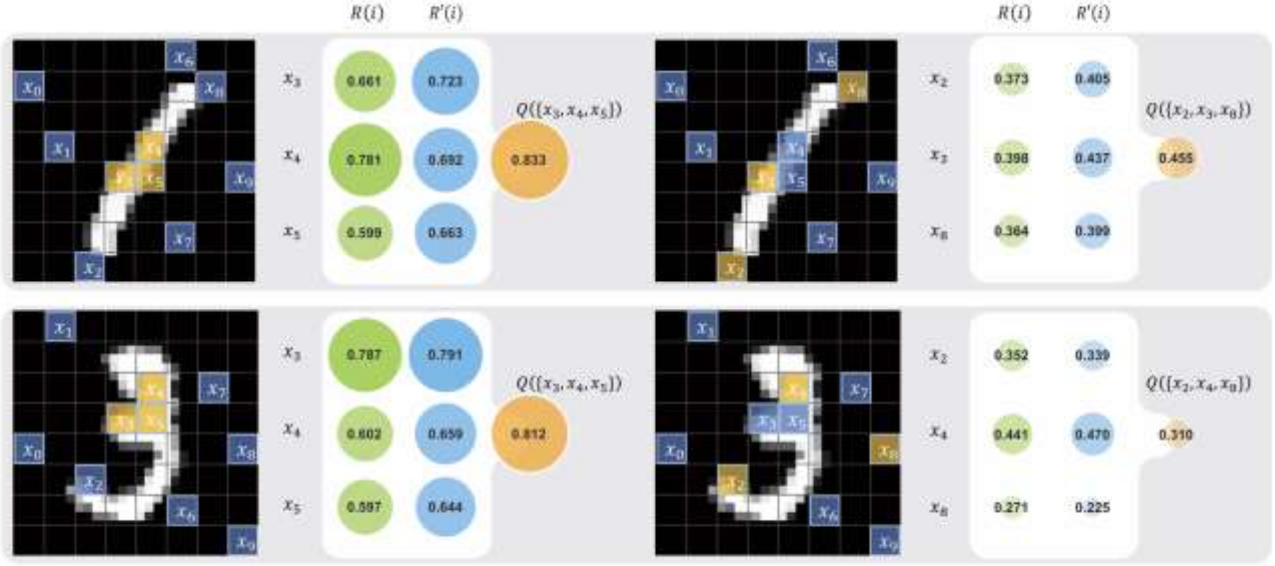

Table 5. Coalition attribution metrics on SST-2 dataset

Sentences	Bert-large
(a) the **mesmerizing performances** of the leads keep the film grounded and keep the audience riveted.	$Q(\{\text{mesmerizing performances}\}) = 0.743$ $R(\{\text{mesmerizing}\}) = 0.690$, $R'(\{\text{mesmerizing}\}) = 0.682$ $R(\{\text{performances}\}) = 0.677$, $R'(\{\text{performances}\}) = 0.685$
(b) one of creepiest, scariest movies to come along in a long, long time, easily **rivaling blair** witch or the others	$Q(\{\text{rivaling blair}\}) = 0.425$ $R(\{\text{rivaling}\}) = 0.145$, $R'(\{\text{rivaling}\}) = 0.391$ $R(\{\text{blair}\}) = 0.250$, $R'(\{\text{blair}\}) = 0.466$
Sentences	**LLaMA**
(a) the **mesmerizing performances** of the leads keep the film grounded and keep the audience riveted.	$Q(\{\text{mesmerizing performances}\}) = 0.746$ $R(\{\text{mesmerizing}\}) = 0.611$, $R'(\{\text{mesmerizing}\}) = 0.652$ $R(\{\text{performances}\}) = 0.726$, $R'(\{\text{performances}\}) = 0.739$
(b) one of creepiest, scariest movies to come along in a long, long time, easily **rivaling blair** witch or the others	$Q(\{\text{rivaling blair}\}) = 0.312$ $R(\{\text{rivaling}\}) = 0.238$, $R'(\{\text{rivaling}\}) = 0.429$ $R(\{\text{blair}\}) = 0.277$, $R'(\{\text{blair}\}) = 0.286$

Application: explaining the Go game

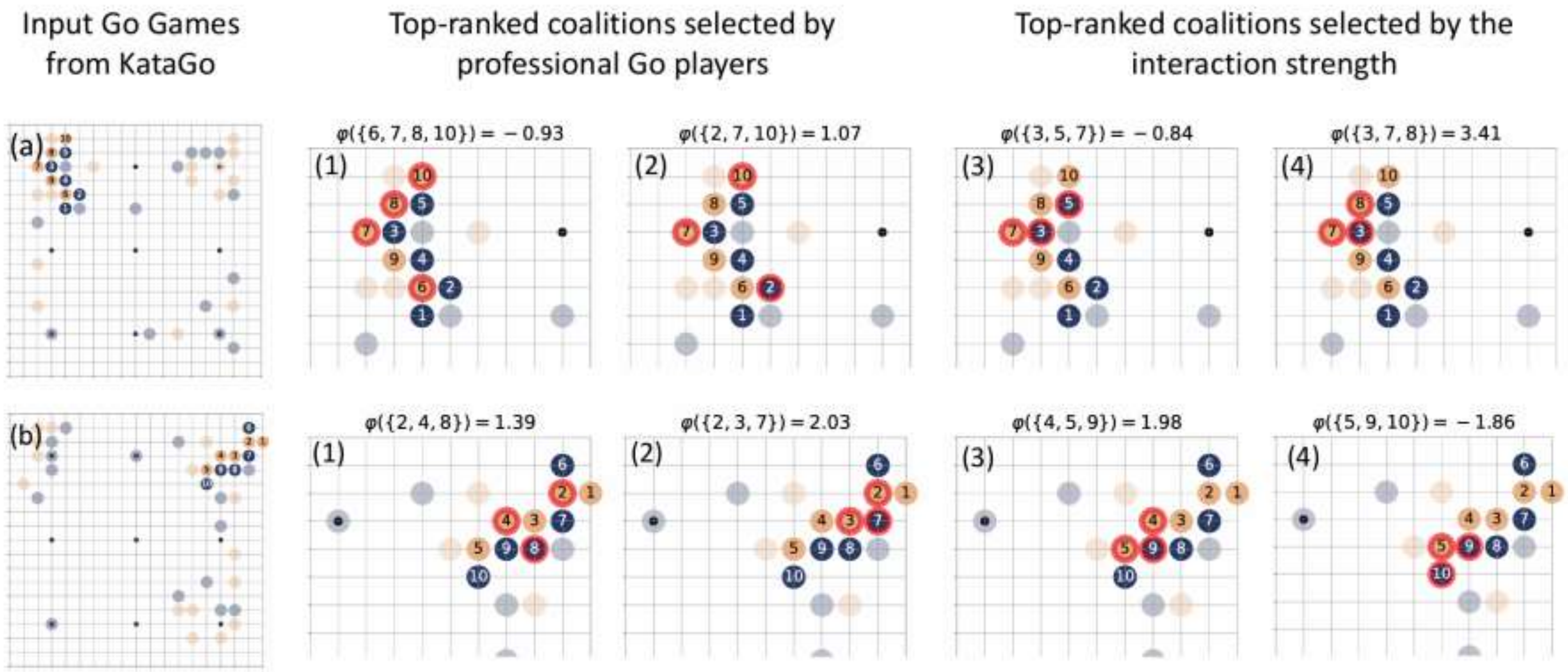

Figure 2. Visualization of two approaches for the selection of coalitions in KataGo. For a coalition S, $\varphi(S) > 0$ means the coalition S of stones makes a positive numerical effect for the white, while it makes a negative effect when $\varphi(S) < 0$.

Thank you!

- Contact:
  - Xinhao Zheng :
    void_zxh@sjtu.edu.cn
  - Prof. Quanshi Zhang
    zqs1022@sjtu.edu.cn