IMPROVING EXPLANATION RELIABILITY THROUGH GROUP ATTRIBUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Although input attribution methods are mainstream in understanding predictions of DNNs for straightforward interpretations, the non-linearity of DNNs often makes the attributed scores unreliable in explaining a given prediction, deteriorating the faithfulness of the explanation. However, the challenge could be mitigated by attributing scores to groups of explanatory components instead of the individuals, termed group attribution. While a group attribution would explain the group-wise contribution more reliably, it does not explain the component-wise contributions so that estimating component-wise scores yields less reliable explanation, indicating the trade-off of group attributions. In this work, we introduce the generalized definition of reliability loss and group attribution, and formulate the optimization problem of the trade-off with these terms. We apply our formalization to Shapley value attribution to propose the optimization method G-SHAP. We show the effectiveness and explanatory benefits of our method through empirical results on image classification tasks.

1 INTRODUCTION

The advance in deep learning facilitates a training model to learn high-level semantic features, but intrinsic difficulties in explaining predictions of DNNs become a primary barrier to real-world applications. While various approaches have been proposed to tackle the challenge, which includes deriving global behavior or knowledge of a trained model (Kim et al., 2018), explaining the semantics of a target neuron in a model, (Ghorbani et al., 2019; Simonyan et al., 2013; Szegedy et al., 2015), introducing self-interpretable models (Zhang et al., 2018; Dosovitskiy et al., 2020; Touvron et al., 2020; Arik & Pfister, 2019), or other explanatory methods, input-attribution methods became mainstream as they explains the contribution of each explanatory component with a scalar score, providing a straightforward post-hoc explanation for a given prediction.

Despite the intuitiveness, however, non-linearity in DNNs would make their attributed scores less reliable in explaining their actual contributions. It results in the discrepancy between the explained and actual model behavior for a prediction, deteriorating the faithfulness of the explanation.

Due to the inherent difficulty in explaining non-linear contributions with scalar scores, a couple of studies have tackled the challenge by explaining higher-order statistics such as interaction: (Grabisch & Roubens, 1999) formalizes the axiomatic interactions for cooperative games, (Tsang et al., 2018) explains the statistical interaction between input features from learned weights in DNN, (Kumar et al., 2021) introduces Shapley Residuals to quantify the unexplained contribution of Shapley values, (Janizek et al., 2021) extends Integrated Gradients (Sundararajan et al., 2017) to Integrated Hessians to explain the interaction between input features.

While those methods better explain non-linear behavior of a given prediction, their explanation would be less intuitive to visualize, limiting the usability of the explanations.

Instead, it can be alleviated by explaining group-wise contributions, termed group attribution, where components in each group are considered as one shared attributable component. Appropriate grouping could reduce unexplained behavior in component-wise contributions in each group, yielding more reliable group scores in explaining their contributions.

However, since component-wise contributions in each group are not explained by group scores, their estimated scores would be less reliable than the ordinary component-wise scores. In other words, a

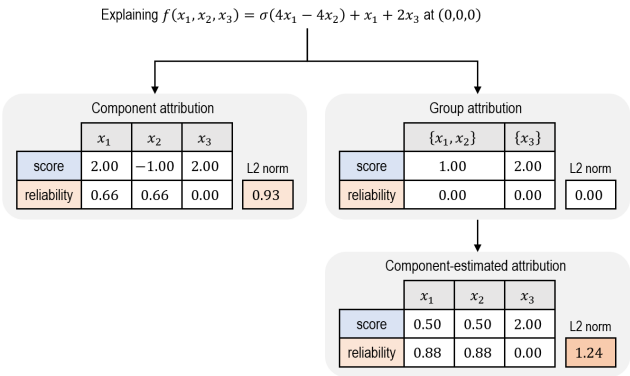

Figure 1: Trade-off of reliability loss of group score and component-estimated score of a group attribution. Grouping x_1, x_2 resolves their component-wise non-linearity so that reliability loss of the group score $\phi_{\{1,2\}}$ decreases but that of estimated component-wise score from the group score increases. Here the metric of attribution score and its reliability loss are defined as $\phi_i(\mathbf{x}) = \frac{\partial}{\partial x_i} f(\mathbf{x})$ and $\xi_i^2(\mathbf{x}) = \mathbb{E}_{t \sim \mathcal{N}(0,1)}[(f(\mathbf{x} + t\mathbf{e}_i) - f(\mathbf{x}) - t\phi_i)^2]$, respectively.

group attribution would be more reliable in explaining group-wise contributions but less reliable in component-wise contributions.

Therefore, reliability loss of both a group attribution and its component-estimated attribution should be considered to find a grouping, implying a trade-off optimization problem. Figure 1 illustrates this problem with simple non-linear function, where the attribution scores and their reliability loss are given as the input gradient and expected L2 error of its tangent approximation, respectively.

Our work consists as follows: First we introduce the generalized definition of reliability loss and group attribution, utilized to formulate the optimization problem for group attributions. We then apply our formalization to Shapley value(Lundberg & Lee, 2017) and propose the optimization method G-SHAP. Finally, we present the empirical results of our method on image classification tasks and show the effectiveness and explanatory benefits.

Our contributions are summarized as follows:

1. We introduce a novel perspective to assess and improve the explainability of attributed explanations, which are **reliability loss** that indicates the discrepancy between the explained and the actual contributions, and **group attribution** that attributes scores to group of explanatory components instead of the individuals, respectively. Since a group attribution would be less reliable in explaining component-wise contributions, we formulate the optimization problem that address the reliability trade-off. Our formulation consists of generalized terms and axioms, applicable for general input attribution methods.

2. We propose **G-SHAP**, the optimization method for Shapley value attribution which utilizes novel statistics of Shapley value. We empirically verify the effectiveness and explanatory benefits of our method through comparison with other grouping methods.

2 General formalization for Reliability loss and Group attribution

2.1 Reliability loss of an attribution

Let $\mathbf{x}^* = (x_1^*, ..., x_N^*) \in \mathbb{R}^N$ be an input data with attributable components $X = \{x_1, ..., x_N\}$ and $f : \mathbb{R}^N \to \mathbb{R}$ be a model function to explain. Then an input attribution method can be defined as a function Φ, which takes a model function and a prediction point and returns corresponding the attribution scores. Since explaining $f(\mathbf{x})$ at $\mathbf{x} = \mathbf{x}^*$ is equivalent to explaining $f(\mathbf{x} + \mathbf{x}^*)$ at $\mathbf{x} = \mathbf{0}$,

we have

$$\boldsymbol{\phi} = (\phi_1, ..., \phi_N) = \Phi(f, \mathbf{x}^*) = \Phi(f^*, \mathbf{0}) \tag{1}$$

where $f^*(\mathbf{x}) = f(\mathbf{x}+\mathbf{x}^*)$. As the attribution scores are derived according to their scoring policy, the reliability in explaining each explanatory component with a given score can be quantified through the reliability loss measure Ξ, which takes a model function, a prediction point, and an given scores $\mathbf{a} = (a_1, ..., a_N)$ and returns the corresponding reliability losses. It can be formulated as

$$\boldsymbol{\xi}(\mathbf{a}) = (\xi_1(a_1), ..., \xi_N(a_N)) = \Xi(f^*, \mathbf{0}, \mathbf{a}) \tag{2}$$

where all the $\xi_i(a_i)$ are non-negative and lower value implies higher reliability of the given score a_i. It is notable that the given scores $\mathbf{a}$ can be arbitrary, not necessarily $\boldsymbol{\phi}$.

We also define the **total reliability loss** by taking the L2 norm of $\boldsymbol{\xi}(\mathbf{a})$ as

$$\xi(\mathbf{a}) = \|\boldsymbol{\xi}(\mathbf{a})\|_2 = \|\Xi(f^*, \mathbf{0}, \mathbf{a})\|_2 \tag{3}$$

where lower $\xi(\mathbf{a})$ implies the that $\mathbf{a}$ better explains the given prediction. Since $\boldsymbol{\phi}$ is the attribution scores of given prediction, it is expected that ξ would decrease when $\mathbf{a}$ get closer to $\boldsymbol{\xi}$ so that it has global minimum at $\mathbf{a} = \boldsymbol{\xi}$.

2.2 Group attribution and group-wise reliability loss

As mentioned in the introduction, a group attribution attributes a score to each group of explanatory components by treating their components as one shared explanatory component. Formally, let $\mathbf{G} = \{G_1, ..., G_M\}$ be a grouping (partition) of the component set $X = \{x_1, ..., x_N\}$. Then the group-mapped function $f^*_{\mathbf{G}}$ assigns each group-variable g_i to its corresponding components variables of X, defined as

$$f^*_{\mathbf{G}}(g_1, ..., g_M) = f^*(g_{\sigma(1)}, ..., g_{\sigma(N)}) \tag{4}$$

where σ is the group map such that $x_i \in G_{\sigma(i)}$ for each $1 \leq i \leq N$.

Then the group attribution $\boldsymbol{\phi}_{\mathbf{G}}$ is defined as the attribution scores of $f^*_{\mathbf{G}}$, which is

$$\boldsymbol{\phi}_{\mathbf{G}} = (\phi_{G_1}, ..., \phi_{G_M}) = \Phi(f^*_{\mathbf{G}}, \mathbf{0}) \tag{5}$$

By the definition, each group score ϕ_{G_j} indicates the aggregated contribution of their components $x_i \in G_j$ but is not necessarily equal to the sum of the component scores in general.

Similarly, we can derive the group-wise reliability loss of explaining each explanatory group G_j with a_{G_j} and their **total group-wise reliability loss (GR)** as below.

$$\boldsymbol{\xi}(\mathbf{a}_{\mathbf{G}}) = (\xi_{G_1}(a_{G_1}), ..., \xi_{G_M}(a_{G_M})) = \Xi(f^*_{\mathbf{G}}, \mathbf{0}, \mathbf{a}_{\mathbf{G}}), \qquad \xi(\mathbf{a}_{\mathbf{G}}) = \|\boldsymbol{\xi}(\mathbf{a}_{\mathbf{G}})\|_2 \tag{6}$$

In this case, however, the reliability loss measure Ξ is required to not have a bias to the group size for L2 norm calculation, or comparing GR among arbitrary groupings would not be feasible.

2.3 Component-wise reliability loss of group attribution

Since a group score does not explain its component-wise contributions, their component-wise scores can only be estimated from the group score. We introduce an **score estimating function** ζ, which takes a grouping and group scores $\mathbf{a}_{\mathbf{G}} = (a_{G_1}, ..., a_{G_M})$ and returns the estimated component-wise scores, given as

$$\tilde{\mathbf{a}_{\mathbf{G}}} = (\tilde{a}_1, ..., \tilde{a}_N) = \zeta(\mathbf{G}, \mathbf{a}_{\mathbf{G}}) \tag{7}$$

where the tilde denotes the estimated component scores.

Once we have the estimated component scores, the component-wise reliability loss and their **total component-wise reliability loss (CR)** are consequently given as below.

$$\boldsymbol{\xi}(\tilde{\mathbf{a}_{\mathbf{G}}}) = (\xi_1(\tilde{a}_1), ..., \xi_N(\tilde{a}_N)) = \Xi(f^*, \mathbf{0}, \tilde{\mathbf{a}_{\mathbf{G}}}), \qquad \xi(\tilde{\mathbf{a}_{\mathbf{G}}}) = \|\boldsymbol{\xi}(\tilde{\mathbf{a}_{\mathbf{G}}})\|_2 \tag{8}$$

While the ζ can be arbitrary defined, it is required not to utilize any information of given prediction, or the estimated score would contain unexpected information of the actual component-wise contributions.

2.4 OPTIMIZING GROUP ATTRIBUTION

Since a desired group attribution is expected to have GR lower (better) than $\xi(\boldsymbol{\phi})$ but have CR higher (worse) than $\xi(\boldsymbol{\phi})$, we need to introduce the total score to optimize their trade-off effect.

For GR, we regard $\xi(\boldsymbol{\phi})$ as baseline so that we define **normalized GR score (NGR)** $\mathcal{G}$ as ratio of the improved amount to the baseline, given as

$$\mathcal{G}(\mathbf{G}) = (\xi(\boldsymbol{\phi}) - \xi(\boldsymbol{\phi}_{\mathbf{G}}))/\xi(\boldsymbol{\phi}) \tag{9}$$

It follows that higher $\mathcal{G}$ implies more reliability of group-wise scores: it becomes 1 if $\xi_{\mathbf{G}} = 0$ (maximum improvement) and 0 if $\xi(\boldsymbol{\phi}_{\mathbf{G}}) = \xi(\boldsymbol{\phi})$ (no improvement). It can be negative if the grouping is ill-chosen.

On the other hand, CR is expected to be higher than the baseline $\xi(\boldsymbol{\phi})$ but does not have trivial upper bound. Hence we regard the $\xi(\tilde{\boldsymbol{\phi}_{\mathbf{G}_{\text{all}}}})$ as the upper bound, where $\mathbf{G}_{\text{all}}$ stands for merging all components into one group, i.e, $\mathbf{G}_{\text{all}} = \{G_1\} = \{\{x_1, ..., x_N\}\}$. Then **normalized CR score (NCR)** as non-deteriorated amount to the bound gap, given as

$$\mathcal{C} = (\xi(\tilde{\boldsymbol{\phi}_{\mathbf{G}_{\text{all}}}}) - \xi(\tilde{\boldsymbol{\phi}_{\mathbf{G}}}))/(\xi(\tilde{\boldsymbol{\phi}_{\mathbf{G}_{\text{all}}}}) - \xi(\boldsymbol{\phi})) \tag{10}$$

It also follows that higher $\mathcal{C}$ implies more reliability of estimated component-wise scores: it becomes 1 if $\xi(\tilde{\boldsymbol{\phi}_{\mathbf{G}}}) = \xi(\boldsymbol{\phi})$ (no deterioration), 0 if $\xi(\tilde{\boldsymbol{\phi}_{\mathbf{G}}}) = \xi(\tilde{\boldsymbol{\phi}_{\mathbf{G}_{\text{all}}}})$ (deteriorated as $\mathbf{G}_{\text{all}}$). It can even be negative if the grouping is extremely ill-chosen.

Since there exist two singular cases that should be avoided, which are singleton grouping (no-grouping) and the all-grouping $\mathbf{G}_{\text{all}}$ with their Once we have the NGR and NCR scores, we define the total score $\mathcal{L}$ as the geometric mean of two scores since two singular cases (no-grouping, all-grouping) yields NGR, NCR scores as $(0, 1)$ and $(1, 0)$, respectively. It is defined as

$$\mathcal{L} = \max\{(\mathcal{G} + \epsilon)/(1 + \epsilon), 0\}^{\frac{1}{2}-\beta} \max\{(\mathcal{C} + \epsilon)/(1 + \epsilon), 0\}^{\frac{1}{2}+\beta} \tag{11}$$

where $\epsilon \geq 0$ is the tolerance hyperparameter for dealing with negative NGR and NCR values and $\beta \in [-1/2, 1/2]$ is the balancing hyperparameter such that positive β weighs more to NCR than NGR and vice versa.

3 APPLICATION TO SHAPLEY VALUE

Shapley value came from cooperative game theory, indicating fair distribution of total reward to each player involved in. It can be utilized as the axiomatic attribution scores for post-hoc model explanations (Lundberg & Lee, 2017), where the players and the reward are corresponded to the binary explanatory components and the output difference of the target model function, respectively.

3.1 RELIABILITY LOSS OF SHAPLEY VALUE

Formally, let $Z = \{z_1, ..., z_N\}$ be the set of explanatory components and $\mathcal{Z} = \{0, 1\}^N$ be the set of the all possible involvement states $\mathbf{z} = (z_1, ..., z_N)$, where each $z_i = 1, 0$ indicates whether z_i is involved or not, respectively. Once the target model function $f : \mathcal{Z} \to \mathbb{R}$ is given, **contribution** of z_i at a state $\mathbf{z}$ is given as

$$h_i(\mathbf{z}) = f(\mathbf{z}_{i=1}) - f(\mathbf{z}) \tag{12}$$

where $z_{i=1}$ denotes the assigned $\mathbf{z}$ with $z_i = 1$. Since $h_i(\mathbf{z})$ is trivially zero when $z_i = 1$, we restrict the domain of h_i as $\mathcal{Z}_{i=0} = \{\mathbf{z} \in \mathcal{Z} | z_i = 0\}$.

Shapley value of z_i is given as weighted sum of all possible contributions h_i, which is

$$\phi_i = \sum_{\mathbf{z} \in \mathcal{Z}_{i=0}} w_N(|\mathbf{z}|) h_i(\mathbf{z}), \qquad w_N(k) = \frac{|k|!(N - |k| - 1)!}{N!} \tag{13}$$

where $|\mathbf{z}|$ denotes the number of ones in $\mathbf{z}$, termed level of $\mathbf{z}$.

It is notable that $\sum_{|\mathbf{z}|=k} w_N(k) = 1/N$, which implies Shapley value is the mean of level-wise mean of h_i. Hence Shapley value ϕ_i can be considered as the expectation of h_i by regarding the

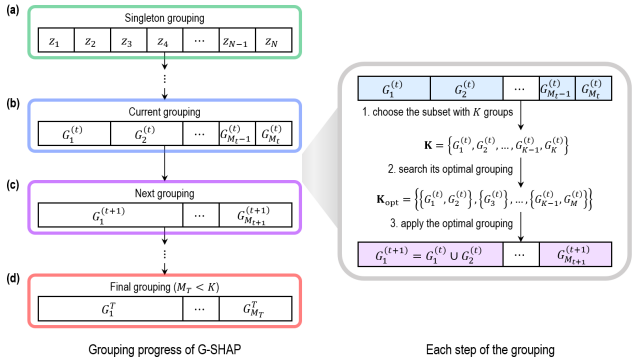

Figure 2: Illustration of the G-SHAP algorithm: The algorithm starts with component-attribution that (NGR, NCR)=(0, 1), continues to apply the optimal grouping of the subset, and finishes with all-grouped attribution. G-SHAP finally takes the best group attribution among the grouping steps. Here ϵ is not considered for simplicity.

weights as the probability, stated as $\phi_i = \mathbb{E}_{w_N}[h_i]$. It leads to define the reliability loss of z_i as the expected L2 error of h_i, which is

$$\xi_i^2(a_i) = \mathbb{E}_{w_N}[(h_i - a_i)^2] \tag{14}$$

Since it follows that $\xi_i^2(a_i) = \xi_i^2(\phi_i) + (a_i - \phi_i)^2$, the total reliability loss is represented as

$$\xi^2(\mathbf{a}) = \xi^2(\boldsymbol{\phi}) + \|\mathbf{a} - \boldsymbol{\phi}\|_2^2 \tag{15}$$

which indicates the total reliability loss has the minimum value $\xi^2(\boldsymbol{\phi})$ when $\mathbf{a} = \boldsymbol{\phi}$, termed **Shapley variance** of z_i.

3.2 G-SHAP: Optimizing group attribution for Shapley value

As a group attribution considers components in each group as a shared explanatory component, marginal contribution of a group $G = \{z_{i_1}, ..., z_{i_M}\} \subseteq Z = \{z_1, ..., z_N\}$ indicates the output difference of f by switching all $z_i \in G$ from 0 to 1, formulated as

$$h_G(\mathbf{z}) = f(\mathbf{z}_{G=1}) - f(\mathbf{z}) = \sum_{1 \le k \le M} h_{i_k}(\mathbf{z}_{G=I_k}) \tag{16}$$

where the subscript $G = I_k$ means $z_{i_1}, ..., z_{i_{k-1}}$ are 1 and others are 0.

As the group contribution consists of M terms of $M-1$ conditioned contribution, evaluating Shapley values and variances of an arbitrary group of M components requires all those $\mathcal{O}(N^(M-1))$ contributions, which is infeasible.

However, a conditioned contribution $h_{i|j=0}$ or $h_{i|j=1}$ can be approximated by the marginal contribution h_i if the z_j has little effect on the h_i, similar to the approach of (Guanchu, 2022). First, marginal Shapley value and variances of z_i can be decomposed into z_j conditioned terms as

$$\begin{aligned}
\phi_i &= \sum_{k=0}^{N-1} \sum_{|\mathbf{z}|=k} w_N(k) h_i(\mathbf{z}) = \sum_{k=0}^{N-2} \sum_{|\mathbf{z}|=k} w_N(k) h_{i|j=0}(\mathbf{z}) + w_N(k+1) h_{i|j=1}(\mathbf{z}) \\
\xi_i^2 &= \sum_{k=0}^{N-1} \sum_{|\mathbf{z}|=k} w_N(k)(h_i(\mathbf{z}) - \phi_i)^2 = \sum_{k=0}^{N-2} \sum_{|\mathbf{z}|=k} w_N(k)(h_{i|j=0}(\mathbf{z}) - \phi_i)^2 \\
&\qquad + w_N(k+1)(h_{i|j=1}(\mathbf{z}) - \phi_i)^2
\end{aligned} \tag{17}$$

Since the weight summation property $w_{N-1}(k) = w_N(k) + w_N(k+1)$ holds, it implies that if $h_{i|j=0}$ and $h_{i|j=1}$ converges to the marginal contribution h_i then the conditioned Shapley value

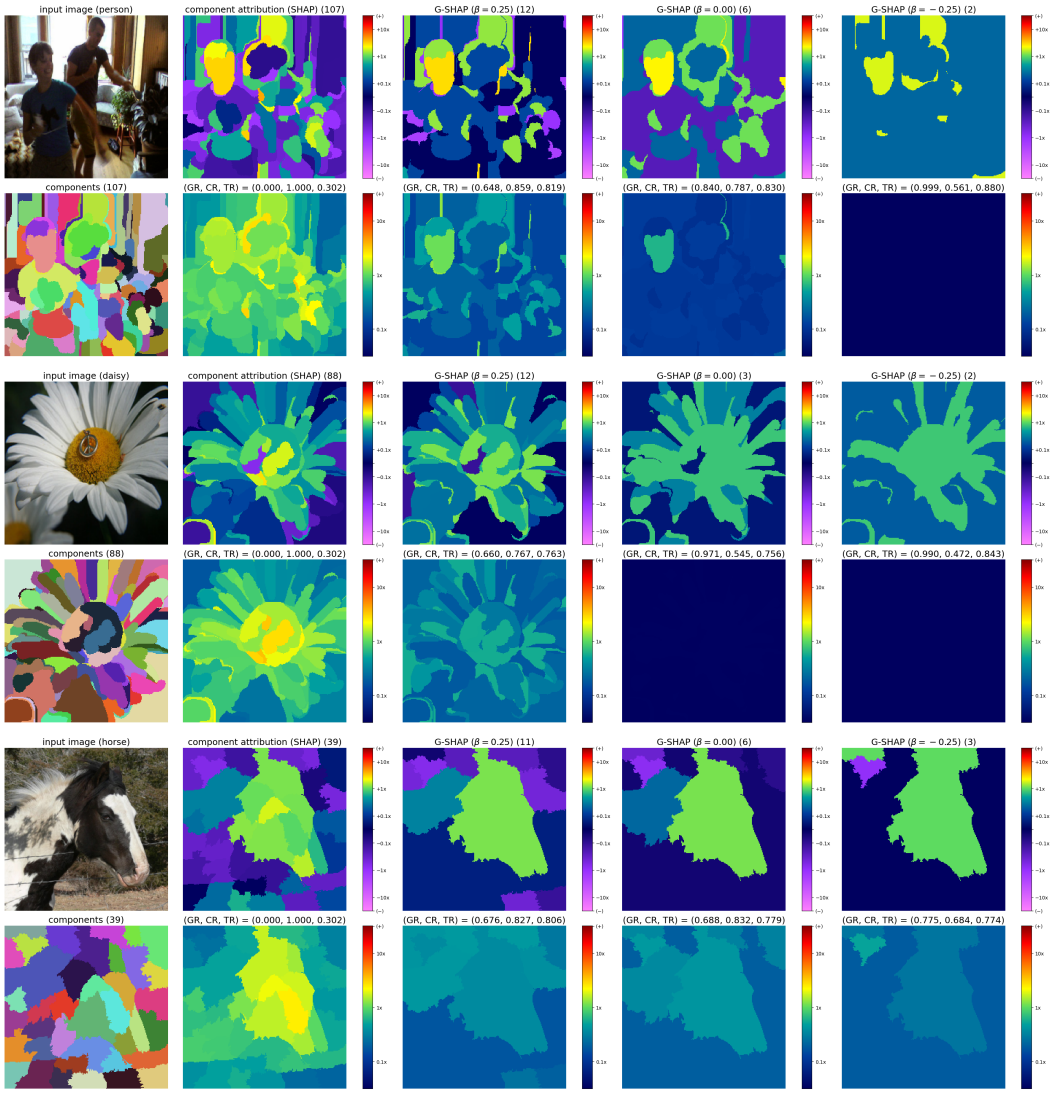

Figure 3: G-SHAP results for the image classification task, taken from MS COCO, Flower5, Pascal VOC dataset, where superpixels are chosen as graph-based, graph-based, and quick-shift, for each image, respectively. The 3-5th columns stand for $\beta = 0.25, 0.00$ and -0.25, respectively. For each image, heatmap of the upper row indicates the attribution score and the lower row indicates the attribution reliability. Heatmaps are area-normalized ratio to their base values, which are their sum divided by entire area of the image. Here the GR, CR, and TR score indicates the NGR, NCR, and the total score $\mathcal{L}$, respectively. More experimental results and codes are available at `https://anonymous.4open.science/r/G_SHAP-5F4E/README.md`

$\phi_{i|j=0}, \phi_{i|j=1}$ and variances $\xi^2_{i|j=0}, \xi^2_{i|j=1}$ also converges to the marginal terms ϕ_i and ξ^2_i, respectively. This implies that Shapley value and variances of group G in X can be estimated at those in $X \setminus J$, where $J \subseteq X \setminus G$ is the set of neglected components such that $h_{i|j=0}$ and $h_{i|j=1}$ to close h_i for all $z_i \in G$ and $j \in J$, respectively.

In our method G-SHAP, we take $\epsilon_i = \sum_{j \neq i} (\phi_{i|j=0} - \phi_i)^2 + (\phi_{i|j=1} - \phi_i)^2$ as the heuristic for choosing the core searching set $K = X \setminus J$. Then we search the optimal grouping in K and apply the grouping to the Z. After the grouping is applied, each group is considered as new explanatory component and the progress iterates until it ends with $k < |K|$ groups. The overall progress is illustrated in the Figure 2.

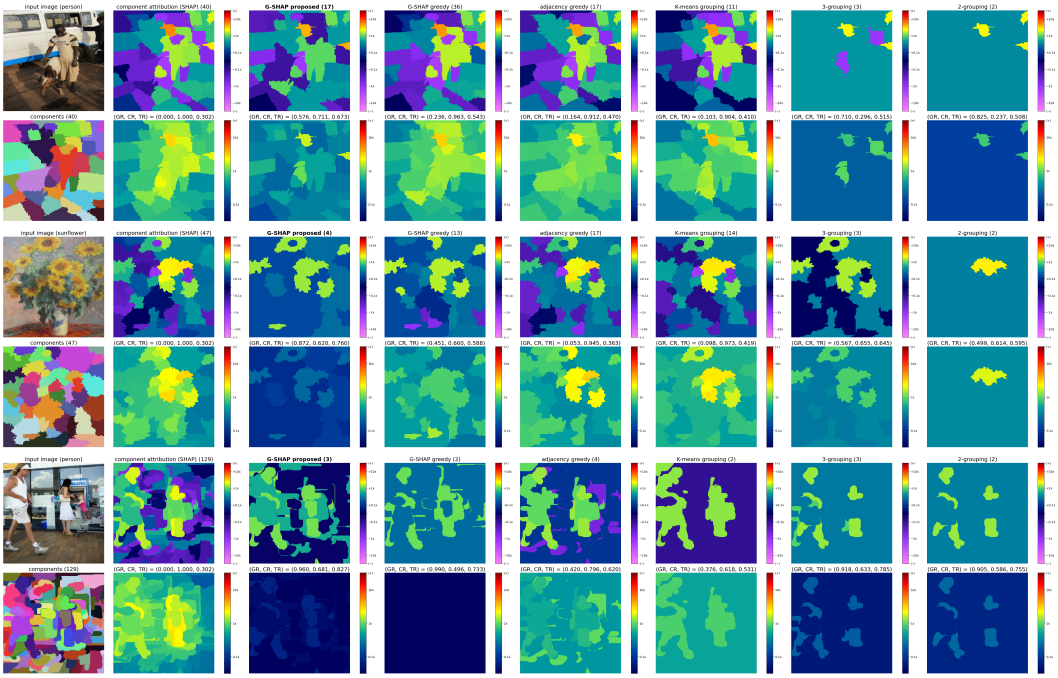

Figure 4: Comparison results of G-SHAP with various heuristic methods, where the images are taken from COCO, Flower5, and Pascal VOC dataset, and the superpixels are chosen from quick-shift, quick-shift, and graph-based method, respectively.

4 Experimental Results

We have experimented our method for image classification tasks to verify the explanatory benefits of group attribution. Since our research goal is to improve reliability of attribution scores, typical assessment metrics such as measuring AUC in deletion game Petsiuk et al. (2018); Wagner et al. (2019) or application specific metrics such as localization ability would not validate our results. Instead, we first show optimization effect of our method through quantitative and qualitative analysis. Second, we perform abliation studies to show effectiveness of our grouping policy by comparing with baseline methods. Finally, we show improved local explainability of our method by estimation game, which utilizes deletion game to measure the error of model output changes.

4.1 Experimental Setup

We have applied the proposed method on the validation datasets of Flower5(multi-class)(Mamaev, 2018), MS COCO 2014(multi-label)(Lin et al., 2014), and Pascal VOC 2012(multi-label)(Everingham et al., 2012) with ImageNet 2012(Russakovsky et al., 2015) pretrained ResNet-50(He et al., 2016) model, where Flower5 stands for subset of Flower dataset with 5 distinctive classes (daisy, dandelion, rose, sunflower, tulip). We have fine-tuned the model to each datasets as the average prediction accuracy is 94.7%, 82.9%, and 90.3%, respectively. For MS COCO and Pascal VOC models, we have taken logit value of top-1 label as the output.

Since our heuristic of choosing the core set requires conditional Shapley values which have $\prime(N^2)$ terms, we have considered superpixels of a image as explanatory components. We have experimented with two superpixel methods, quick-shift(Vedaldi & Soatto, 2008) and graph-based(Felzenszwalb & Huttenlocher, 2004) segmentation method, which existing attribution methods LIME(Ribeiro et al., 2016) and XRAI (Kapishnikov et al., 2019) use. As Shapley value considers binary input states, we have defined the input map ψ as mean-color masking function, and score estimating function ζ as distributing a group score according to pixel area. We set core set dimension $|K| = 10$, $\beta = 0.0$, and $\epsilon = 0.1$ as default.

Superpixel method	Datasets	$\beta = 0.25$		$\beta = 0.00$		$\beta = -0.25$	
		NGR	NCR	NGR	NCR	NGR	NCR
Quick-shift	COCO	0.596	0.832	0.799	0.679	0.939	0.667
	Flower5	0.593	0.831	0.758	0.642	0.919	0.616
	VOC	0.602	0.829	0.824	0.670	0.942	0.680
Graph-based	COCO	0.704	0.779	0.892	0.630	0.973	0.692
	Flower5	0.603	0.779	0.840	0.605	0.950	0.653
	VOC	0.719	0.776	0.905	0.626	0.975	0.667

Table 1: Reliability scores of the G-SHAP for $\beta = 0.25, 0.00$ and -0.25

Methods	Quick-Shift		Graph-based	
	NGR	NCR	NGR	NCR
2-grouping	0.786	0.552	0.818	0.558
3-grouping	0.696	0.646	0.754	0.635
K-means grouping	0.476	0.701	0.493	0.646
Adjacency greedy	0.443	0.720	0.423	0.711
G-SHAP greedy	0.678	0.584	0.648	0.561
G-SHAP porposed	0.794	0.664	0.879	0.620

Table 2: Comparison with baseline heuristics, where scores are averaged on the datasets

4.2 Optimization effects of G-SHAP

We have observed NGR, and NCR to verify the improved and non-deteriorated reliability loss of G-SHAP attribution for $\beta = -0.25, 0.00, 0.25$, stated at Table 1. As NGR, NCR indicates normalized ratio of the amounts, it tells that $75\% \sim 91\%$ of baseline reliability $\xi(\boldsymbol{\phi})$ are resolved through grouping while $62\% \sim 68\%$ of reliability gap $\xi(\tilde{\boldsymbol{\phi}}_{\mathbf{G}_{\text{all}}}) - \xi(\boldsymbol{\phi})$ is saved for $\beta = 0.00$ case. It has also been observed that the NGR and NCR value are considerably affected by β such that positive β weighs NCR much than NGR, whereas the negative β weighs NCR much than NGR, which meets to our expectation.

Figure 3 tells that the balancing effect of the β can be confirmed in qualitatively, as the higher β results in higher NCR that the heatmap of G-SHAP is closer to the component attribution (SHAP) but lower NGR that group-wise reliability losses are less improved. Lower β also results in the opposite as well but also tells that G-SHAP attribution consists of few groups with salient superpixels. This have motivated to compare our method with heuristic grouping methods that yields fewer groups or merging component/groups with close attribution scores, discussed in the later subsection.

4.3 Comparison and Ablation studies

In order to show the validity of our grouping strategy, we have compared G-SHAP with various grouping policies, stated in the Table 2 and illustrated in the Figure 4. We have described the selected heuristic methods as follows.

First, we have employed **2-grouping** and **3-grouping** methods to confirm the performance of naive grouping. The 2-grouping method sort the components by attribution scores and split into two groups by merging top $1 \leq k \leq N$ components and the others, and returns the best one. Similarly, the 3-grouping method considers all cases of top-k and bottom-m and the middle-$(N - k - m)$ grouping and pick the best one. While the results show that their NGR and NCR are slightly lower than ours (less than 0.1), they show very limited information for the prediction as only the most or few salient superpixels are explained.

We have also employed **K-means grouping** and **Adjacency-greedy grouping** methods to test the grouping performance of merging component with closer attribution scores, tended to save NCR. The K-means grouping method performs clustering of the components with $2 \leq k \leq 10$ clusters and returns the best grouping, where the distance metric is given as difference of normalized Shapley value (divided by superpixel area). The **Adjacency-greedy** method iteratively merge two groups

		min-deletion			max-deletion			random-deletion		
Superixel method	Attribution	Flower5	COCO	VOC	Flower5	COCO	VOC	Flower5	COCO	VOC
Quick-shift	SHAP	0.476	0.566	0.671	0.444	0.438	0.515	0.446	0.465	0.587
	G-SHAP	0.171	0.158	0.204	0.159	0.144	0.191	0.147	0.131	0.169
Graph-based	SHAP	0.532	0.674	0.786	0.458	0.494	0.563	0.469	0.537	0.646
	G-SHAP	0.223	0.176	0.223	0.191	0.161	0.190	0.177	0.149	0.180

Table 3: Estimation game result of component attribution (SHAP) and group attribution (G-SHAP)

with the closest normalized Shapley values, and returns the best grouping. Unlikely to the above methods, it shows that NGR is clearly lower than our method, whereas NCR is similar or slightly higher than our methods. It implies that these two method have preserved NCR as expected but failed to resolve NGR as any interaction statistics are utilized.

Finally, we have experimented ablation study of G-SHAP without core set searching, which instead greedily merge two groups which are expected to improve the $\mathcal{L}$ the most, named **G-SHAP greedy**. As both NGR and NCR are around 0.1 lower than our method in average, this implies that the optimization problem is challenging to solve with simple greedy approach so that optimal partition searching in the core set is necessary.

4.4 Estimation games

Deletion game Petsiuk et al. (2018); Wagner et al. (2019) is main strategy to assess the attribution scores which removes each component of input data in sequence and evaluate the model output drop through AUC of the curve. For example, max-deletion game removes components in decreasing order of attribution score so that low AUC value is expected. However, these methods generally take the ranking of the attribution score only so that they would not assess the reliability information of the scores. Therefore, we have employed this idea in a different way, termed **estimation game** which aims to measure error of output changes of target model under deletion process. As Shapley value indicates expected output changes, this assessment approach has been also utilized in (Guanchu, 2022). We have employed three types of deletions: min-deletion, max-deletion, and random-deletion, which deletes inputs in increasing, decreasing, and random order of attribution score, respectively. Since output value can be arbitrary scaled depending on the prediction input, we have normalized as follows: (1) we have linearly rescaled y-axis such that $y = 0, 1$ stands for ground image (mean-colored image) and target image, respectively. (2) we have also linearly rescaled x-axis as it indicates ratio of removed pixels to entire pixels. Therefore removal game always start from $(0, 1)$ and ends with $(1, 0)$ after the normalization. As we can see the result in Table 3, G-SHAP resolves around 60% to 70% of L2 estimation error of component attribution (SHAP), implying that G-SHAP attribution provides better understanding of local behavior of the model.

5 Conclusion

As understanding and reasoning prediction of machine learning model become more demanding for both in application needs and academical research, various methods have been proposed to address the problem and provided more insights of the behavior of model. Though input-attribution methods provide clear interpretation as the explanations correspond to the data, non-linearity of deep models intrinsically hinders reliability of attribution. In this work, we have presented novel perspective of attributing groups, formulate it with the optimization problem, and introduce G-SHAP, optimization algorithm for Shapley value attribution. In the experiments, our method have shown that improvement of the group-wise reliability loss is clearly larger than deterioration of component-wise reliability loss, indicating the explanatory benefit. It implies that the prediction can be better understood with the groups than with individual components, suggesting that each group is key feature for the prediction and their contribution to the model is disentangled. Moreover, it shows the potential of localizing abaility as the salient components are merged into few groups with corresponding semantics. While we have verified the explanatory benefits of our method both in qualitative and quantitative way, our method utilizes Shapley conditional terms and search partition spaces with iteration, which results high complexity to start with pixel-wise components. Deeper analytical approach would improve the performance and feasibility, left as potential for future works.

REFERENCES

Sercan O Arik and Tomas Pfister. Tabnet: Attentive interpretable tabular learning. *arXiv preprint arXiv:1908.07442*, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html, 2012.

Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International journal of computer vision*, 59(2):167–181, 2004.

Amirata Ghorbani, James Wexler, James Zou, and Been Kim. Towards automatic concept-based explanations. *arXiv preprint arXiv:1902.03129*, 2019.

Michel Grabisch and Marc Roubens. An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of game theory*, 28(4):547–565, 1999.

Yu-Neng Guanchu. Accelerating shapley explanation via contributive cooperator selection. In *International conference on machine learning*. PMLR, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Joseph D Janizek, Pascal Sturmfels, and Su-In Lee. Explaining explanations: Axiomatic feature interactions for deep networks. *J. Mach. Learn. Res.*, 22:104–1, 2021.

Andrei Kapishnikov, Tolga Bolukbasi, Fernanda Viégas, and Michael Terry. Xrai: Better attributions through regions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4948–4957, 2019.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pp. 2668–2677. PMLR, 2018.

Indra Kumar, Carlos Scheidegger, Suresh Venkatasubramanian, and Sorelle Friedler. Shapley residuals: Quantifying the limits of the shapley value for explanations. *Advances in Neural Information Processing Systems*, 34, 2021.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll'a r, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. URL `http://arxiv.org/abs/1405.0312`.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

Alexander Mamaev. Flowers recognition. *https://www.kaggle.com/alxmamaev/flowers-recognition*, 2018.

Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pp. 1135–1144, 2016.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, 2015. doi: 10.1109/CVPR.2015.7298594.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. *arXiv preprint arXiv:2012.12877*, 2020.

Michael Tsang, Dehua Cheng, and Yan Liu. Detecting statistical interactions from neural network weights. In *International Conference on Learning Representations*, 2018.

Andrea Vedaldi and Stefano Soatto. Quick shift and kernel methods for mode seeking. In *European conference on computer vision*, pp. 705–718. Springer, 2008.

Jorg Wagner, Jan Mathias Kohler, Tobias Gindele, Leon Hetzel, Jakob Thaddaus Wiedemer, and Sven Behnke. Interpretable and fine-grained visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9097–9107, 2019.

Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. Interpretable convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8827–8836, 2018.