A Unified Taylor Framework for Revisiting Attribution Methods

Huiqi Deng[1,*], **Na Zou**[2], **Mengnan Du**[2], **Weifu Chen**[1,†], **Guocan Feng**[1], **Xia Hu**[2]
[1] Sun Yat-Sen University
[2] Texas A&M University
{denghq7@mail2,mcsfgc@mail,chenwf26@mail}.sysu.edu.cn, {nzou1,dumengnan,xiahu}@tamu.edu

Abstract

Attribution methods have been developed to understand the decision making process of machine learning models, especially deep neural networks, by assigning importance scores to individual features. Existing attribution methods often built upon empirical intuitions and heuristics. There still lacks a general and theoretical framework that not only can unify these attribution methods, but also theoretically reveal their rationales, fidelity, and limitations. To bridge the gap, in this paper, we propose a Taylor attribution framework and reformulate seven mainstream attribution methods into the framework. Based on reformulations, we analyze the attribution methods in terms of rationale, fidelity, and limitation. Moreover, We establish three principles for a good attribution in the Taylor attribution framework, i.e., low approximation error, correct contribution assignment, and unbiased baseline selection. Finally, we empirically validate the Taylor reformulations, and reveal a positive correlation between the attribution performance and the number of principles followed by the attribution method via benchmarking on real-world datasets.

Introduction

Attribution methods have become an effective computational tool in understanding the behavior of machine learning models, especially Deep Neural Networks (DNNs) (Du, Liu, and Hu 2019; Samek et al. 2019). They uncover how machine learning models make a decision by calculating the contribution score of each input feature to the final decision. For example, in image classification, the attribution methods infer the contribution of each pixel to the predicted label for a pre-trained model, and usually create saliency maps to visualize the contributions.

Although several attribution methods (Samek et al. 2020) have been proposed recently, they are based on different heuristics and have very limited theoretical understanding and support. For instance, Occlusion-1 and Occlusion-patch observe the changes of the output induced by adjusting each input pixel or patch (Zeiler and Fergus 2014; Zintgraf et al. 2017); Layer-wise Relevance Propagation (LRP) evaluates the contributions of each input to a non-linear neuron according to the corresponding linear weights in the pre-trained model (Bach et al. 2015). The interpretations generated by those attribution methods with such intuitive rationales are difficult to compare and can not be fully trusted. Hence, it's highly desirable to conduct a comprehensive investigation on the rationales, fidelity, and limitations of those various heuristic methods. Specifically, the following important questions need theoretical investigation: **Rationale**–*What model behaviors do these attribution methods actually reveal*; **Fidelity**–*How much can decision making process be attributed*; **Limitations**–*Where they may fail*.

While some attempts have been made to partially answer the questions by unifying a certain kind of attribution methods, such as additive feature attribution (Lundberg and Lee 2017), multiplying a modified gradient with input (Ancona et al. 2018), or first-order Taylor expansion (Samek et al. 2020), the problems are still not addressed well due to two challenges. The first challenge (**Ch1**) stems from the fact that it is difficult to unify most of existing attribution methods, i.e., to reformulate these attribution methods into one framework, because the methods are based on various heuristics, as we discussed above. The second challenge (**Ch2**) is lacking a theoretical attribution framework, which could offer a good description to the attribution problem so as to theoretically reveal the rationale, fidelity, and limitations of the attribution methods.

In this paper, we study the attribution of DNNs to answer the aforementioned three questions by proposing a general Taylor attribution framework and unifying seven mainstream attribution methods into the framework. The basic idea behind the proposed framework is to attribute an approximation function of DNNs, instead of DNNs themselves. The proposed Taylor attribution framework has three features: (1) the framework is based on Taylor expansion, which is able to approximate sufficiently the behavior of black-box DNNs and has a theoretical guarantee on approximation error; (2) Taylor expansion is a polynomial function, in which attribution and analysis become very easy and intuitive; (3) The Taylor attribution framework is very general, it can unify many attribution methods that are to analyze the output change between input sample and baseline.

We then reformulate seven mainstream attribution methods into the proposed Taylor attribution framework by theoretical derivations. The unified reformulations enable us to

[*]This work is done during her visit at Texas A&M University.
[†]Corresponding author

Symbol	Description
$f(\boldsymbol{x})$	DNN model with input $\boldsymbol{x}$
g_K	K-order Taylor expansion function of f
$\epsilon(\tilde{\boldsymbol{x}}, K)$	Taylor approximation error of g_K at $\tilde{\boldsymbol{x}}$
$H_{\boldsymbol{x}}$	Hessian matrix at $\boldsymbol{x}$
$H^d_{\boldsymbol{x}}, H^t_{\boldsymbol{x}}$	Hessian independent and interactive matrix
$T^\alpha, T^\beta, T^\gamma$	Taylor first, second, and high-order terms
$T^{\beta_d}, T^{\gamma_d}$	Second, high-order independent terms
$T^{\beta_t}, T^{\gamma_t}$	Second, high-order interactive terms
a_i	Attribution of feature x_i
a_i^α	Attribution of feature x_i from T^α
$a_i^{\beta_d}, a_i^{\beta_t}$	Attribution of feature x_i from T^{β_d}, T^{β_t}
$a_i^{\gamma_d}, a_i^{\gamma_t}$	Attribution of feature x_i from $T^{\gamma_d}, T^{\gamma_t}$

Table 1: Symbol descriptions in this paper.

examine rationales, measure fidelity, and reveal limitation for the existing attribution methods in a systematic and theoretical way. We analyze the seven attribution methods by their reformulations. Based on the reformulations and analysis, we establish and advocate three principles for a good attribution in the Taylor attribution framework, which are low approximation error, correct contribution assignment, and unbiased baseline selection.

Finally, we empirically validate the proposed Taylor reformulations by comparing the attribution results obtained by the original attribution methods and their Taylor reformulations. The experimental results on MNIST show the two attribution results are almost consistent. We also reveal a strong positive correlation between the attribution performance and the number of principles followed by the attribution method via benchmarking on MNIST and Imagenet. In summary, this paper has three main contributions:

- We propose a general Taylor attribution framework, and theoretically reformulate seven mainstream attribution methods into the attribution framework.
- We analyze the attribution methods by their reformulation in terms of rationale, fidelity, and limitation, and accordingly establish three principles for a good attribution.
- We empirically validate the Taylor reformulations, and reveal the relationship between attribution performance and the three principles on MNIST and Imagenet.

A General Taylor Attribution Framework

In this section, we propose a Taylor attribution framework to understand the decision making process of DNNs. Specifically, given a pre-trained DNN model f and an input sample $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]^{\mathrm{T}} \in \mathbb{R}^n$, the framework aims to infer the contribution of each feature x_i to the prediction $f(\boldsymbol{x})$. We employ a Taylor expansion function g to approximate the DNN model f, and then conduct the attribution in g because g is a polynomial function and easy to attribute.

The Taylor expansion of f expanded at sample $\boldsymbol{x}$ is[1],

$$f(\tilde{\boldsymbol{x}}) - f(\boldsymbol{x}) = g_K(\boldsymbol{x}, \Delta) + \epsilon(\tilde{\boldsymbol{x}}, K),$$

where $g_K(\boldsymbol{x}, \Delta)$ is the K-order Taylor expansion function of f, $\tilde{\boldsymbol{x}}$ denotes a baseline point which acts as a "refernce" state, vector $\Delta := \tilde{\boldsymbol{x}} - \boldsymbol{x}$, and $\epsilon(\tilde{\boldsymbol{x}}, K)$ is the approximation error between $f(\boldsymbol{x})$ and $g_K(\boldsymbol{x}, \Delta)$ at point $\tilde{\boldsymbol{x}}$. The left side of equation, $f(\tilde{\boldsymbol{x}}) - f(\boldsymbol{x})$, represents the output change, which can be considered as the effect of input change Δ. The attribution problem becomes to decompose the effect to each $\Delta_i := \tilde{x}_i - x_i$, the change of feature i. It's difficult to decompose directly the effect due to the complexity of f. As $g_K(\boldsymbol{x}, \Delta)$ is an approximation of the output change, we instead decompose $g_K(\boldsymbol{x}, \Delta)$ into a attribution vector $\boldsymbol{a} = [a_1, \ldots, a_n]^{\mathrm{T}}$, where a_i denotes the attribution score of feature x_i. An overview of Taylor attribution framework is illustrated in Figure 1. For convenience, we summary the descriptions of main symbols in this paper in Table 1.

First-order Taylor Attribution

The first-order Taylor expansion function $g_1(\boldsymbol{x}, \Delta)$ is

$$g_1(\boldsymbol{x}, \Delta) = f_{\boldsymbol{x}}^{\mathrm{T}} \Delta = \textstyle\sum_i f_{x_i} \Delta_i,$$

where f_{x_i} denotes the derivative of f with respect to x_i. The linear approximation function in first-order Taylor expansion, $g_1(\boldsymbol{x}, \Delta)$, is additive across features and can be easily decomposed. It is obvious that $f_{x_i}\Delta_i$ quantifies the contribution of feature x_i, i.e.,

$$a_i = f_{x_i} \Delta_i.$$

Second-order Taylor Attribution

The second-order Taylor expansion has a smaller approximation error ϵ than the first-order one, so that it is expected more faithful to the model f. The second-order Taylor expansion function $g_2(\boldsymbol{x}, \Delta)$ is given by

$$g_2(\boldsymbol{x}, \Delta) = \underbrace{f_{\boldsymbol{x}}^{\mathrm{T}} \Delta}_{T^\alpha} + \underbrace{\frac{1}{2} \Delta^T H_{\boldsymbol{x}} \Delta}_{T^\beta},$$

where $H_{\boldsymbol{x}}$ is the Hessian matrix, i.e., second-order partial derivative matrix, of f at $\boldsymbol{x}$. We denote the first-order and second-order Taylor terms as T^α and T^β, respectively.

The second-order Taylor expansion function $g_2(\boldsymbol{x}, \Delta)$ is indistinct in determining feature contributions compared with first-order one due to the Hessian matrix. To make the attribution more clear, we decompose $H_{\boldsymbol{x}}$ into two matrices, an independent matrix $H^d_{\boldsymbol{x}}$ and an interactive matrix $H^t_{\boldsymbol{x}} := H_{\boldsymbol{x}} - H^d_{\boldsymbol{x}}$. Here $H^d_{\boldsymbol{x}}$ is a diagonal matrix composed of the diagonal elements in $H_{\boldsymbol{x}}$, which describes the second-order isolated effect of features, and $H^t_{\boldsymbol{x}}$ represents the interactive effect between features. $g_2(\boldsymbol{x}, \Delta)$ could be rewritten as the sum of first order terms T^α, second-order independent terms T^{β_d}, and second-order interactive terms T^{β_t},

$$g_2(\boldsymbol{x}, \Delta) = \underbrace{f_{\boldsymbol{x}} \Delta}_{T^\alpha} + \underbrace{\frac{1}{2} \Delta^T H^d_{\boldsymbol{x}} \Delta}_{T^{\beta_d}} + \underbrace{\frac{1}{2} \Delta^T H^t_{\boldsymbol{x}} \Delta}_{T^{\beta_t}}.$$

Accordingly, the attribution of $g_2(\boldsymbol{x}, \Delta)$ to x_i should be

[1] Noted that although the deep relu network is not differentiable such that Taylor expansion is not applicable, we can use networks with softplus activation (approximation of relu) to provide an insight to the rationale behind relu net.

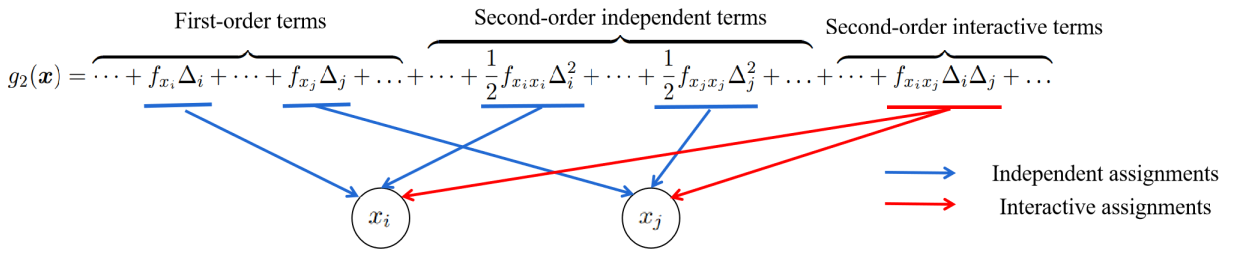

Figure 1: An overview of Taylor attribution framework. Taking a second-order Taylor expansion as an example, $g_2(\boldsymbol{x})$ is composed of first-order, second-order independent, and second-order interactive terms. The first-order and second-order independent terms of x_i can be clearly assigned to x_i, as shown in blue line. The second-order interactive term between x_i and x_j should be and only be assigned to x_i and x_j, as shown in red line.

$$a_i = a_i^{\alpha} + a_i^{\beta_d} + a_i^{\beta_t},$$

where $a_i^{\alpha}, a_i^{\beta_d}$ and $a_i^{\beta_t}$ represent the assigned contributions from T^{α}, T^{β_d} and T^{β_t}, respectively. The contributions from independent terms T^{α} and T^{β_d} can be clearly identified as

$$a_i^{\alpha} = T_i^{\alpha} = f_{x_i}\Delta_i, \quad a_i^{\beta_d} = T_i^{\beta_d} = \tfrac{1}{2}f_{x_ix_i}\Delta_i^2,$$

where T_i^{α} and $T_i^{\beta_d}$ denote the first-order terms and second-order independent terms of feature x_i, respectively.

The difficulty lies on how to assign the contribution from interactive terms T^{β_t}. We propose to handle it by following an intuition behind: the attribution $a_i^{\beta_t}$ is from T^{β_t} and should be the sum of assignments from each interactive effect involving feature x_i,

$$a_i^{\beta_t} = \textstyle\sum_{j\neq i} a_{i,\{x_i,x_j\}}^{\beta_t} = \sum_{j\neq i} w_i^{\{i,j\}} T_{\{x_i,x_j\}}^{\beta_t},$$

where $T_{\{x_i,x_j\}}^{\beta_t} = f_{x_ix_j}\Delta_i\Delta_j$ denotes the second-order interactive terms corresponding to feature x_i and x_j, weight $w_i^{\{i,j\}}$ characterizes the assignment of the interactive terms to x_i, and $a_{i,\{x_i,x_j\}}^{\beta_t}$ is the attribution from $T_{\{x_i,x_j\}}^{\beta_t}$.

The determination of the assignment weight $w_i^{\{i,j\}}$ is complicated and depends on specific case. However, it's considered that *the interactive terms of two features should be only attributed to these two features*. Consider the interactive terms between x_i and x_j, the assignment should satisfy $a_{i,\{x_i,x_j\}}^{\beta_t} + a_{j,\{x_i,x_j\}}^{\beta_t} = T_{\{x_i,x_j\}}^{\beta_t}$, i.e., $w_i^{\{i,j\}} + w_j^{\{i,j\}} = 1$. The second-order interactive term are equally assigned to x_i and x_j, i.e., $w_i^{\{i,j\}} = w_j^{\{i,j\}} = \frac{1}{2}$ in Integrated Gradients (Sundararajan, Taly, and Yan 2017), as shown in the reformulation in Section 3.1.

High-order Taylor Attribution

The analysis on second-order expansion can be naturally extended to high-order expansion where $K > 2$. Let T^{γ} denote all high-order expansion terms, including second-order expansion terms. The high-order Taylor expansion function is

$$g_K(\boldsymbol{x}, \Delta) = T^{\alpha} + T^{\gamma_d} + T^{\gamma_t},$$

where T^{γ_d} and T^{γ_t} denote high-order independent and interactive terms, respectively.

Analogously to the second-order case, the attribution of feature x_i in high-order expansion is given by

$$a_i = a_i^{\alpha} + a_i^{\gamma_d} + a_i^{\gamma_t},$$

where $a_i^{\gamma_d}$, $a_i^{\gamma_t}$ represent the assigned contributions from T^{γ_d} and T^{γ_t}, respectively. The attribution from first-order term and high-order independent term is clear,

$$a_i^{\alpha} = T_i^{\alpha}, \quad a_i^{\gamma_d} = T_i^{\gamma_d},$$

where $T_i^{\gamma_d}$ represent the high-order independent terms of feature x_i. The attribution from interactive terms, $a_i^{\gamma_t}$, consists of all assignments from interactive terms involving x_i,

$$a_i^{\gamma_t} = \textstyle\sum_A a_{i,A}^{\gamma_t}, x_i \in A,$$

where $a_{i,A}^{\gamma_t}$ denotes the attribution from interactive terms corresponding to features in the feature subset A. Note that interactive terms $T_A^{\gamma_t}$ should be only assigned to the features in the subset A, i.e., $\sum_{i\in A} a_{i,A}^{\gamma_t} = T_A^{\gamma_t}$.

The Selection of Baseline Point

From the Taylor attribution framework, the attribution of feature x_i could be seen as a polynomial function of Δ_i (i.e., $\tilde{x}_i - x_i$), and hence it highly depends on $\tilde{x}_i$. Given a baseline of constant vector $\tilde{\boldsymbol{x}} = \boldsymbol{c}$ as many attribution methods did, the attribution of feature whose value is far from c may be overestimated due to a large Δ_i, while the attribution of feature whose value is close to c may be underestimated even if it is important to the decision making process. Such different attributions are a bias in many tasks. For example, in image classification, it's unreasonable to attribute according to the value of features (i.e., pixel values). Specifically, given a black image as baseline, pixels in white color have a large Δ_i close to 255, while pixels in black color have a small Δ_i close to 0. Correspondingly, the attribution methods will biasedly highlight white pixels while neglecting black pixels even if black pixels make up the object of interest. Hence the selection of baseline point $\tilde{\boldsymbol{x}}$ plays a significant role.

Baseline point is used to represent an "absence" of a feature, by which the attribution methods calculate how much the output of the model would decrease considering the absence of the feature (Sturmfels, Lundberg, and Lee 2020). To avoid incorporating aforementioned bias into the attribution process, attribution methods should choose a unbiased baseline which satisfies there is no big differences among Δ_i of different features. That is, Δ_i should be similar to Δ_j for random two feature dimensions. One option is setting Δ as a constant vector $\boldsymbol{c}$, and its corresponding baseline is

$\tilde{\boldsymbol{x}} = \boldsymbol{x} + \boldsymbol{c}$. Such baselines indeed solve the bias issue, however they usually don't make a difference to the output of the model. Another alternative is to neutralize the bias by averaging multiple baselines whose Δs are sampled from some distributions with small variance, such as uniform and Gaussian distributions (Smilkov et al. 2017). Noted that it's difficult for neutralization over samples from a distribution with large variance. This may explain why SmoothGrad and Integrated Gradients will success with small Gaussian noise level while fail with large noise level.

Revisiting Existing Attribution Methods

The proposed Taylor attribution framework is very general to unify most existing attribution methods. On one hand, most mainstream attribution methods aim to assign or decompose the output difference between the input sample and the baseline, $\Delta f = f(\boldsymbol{x}) - f(\tilde{\boldsymbol{x}})$, to each input feature, which can be thought of a function of Δf. On the other hand, the Taylor expansion adopted in the framework could decompose Δf into the sum of input features' effects (Taylor terms). Therefore, the attribution can be unified into our framework, i.e., the attribution could be reformulated as a function of the Taylor terms.

In the following, we first reformulate seven mainstream attribution methods into the proposed framework, and then systematically analyze their rationale, fidelity, and limitations based on the reformulations. Finally, we establish three principles for a good attribution.

Unified Reformulations

We reformulate seven mainstream attribution methods into the proposed Taylor attribution framework, which are Gradient*Input (Shrikumar et al. 2016), Occlusion-1 (Zeiler and Fergus 2014), Occlusion-patch (Zintgraf et al. 2017), DeepLIFT (Rescale) (Shrikumar, Greenside, and Kundaje 2017), ϵ-LRP (Bach et al. 2015), Integrated Gradients (Sundararajan, Taly, and Yan 2017) and Expected Gradients (Erion et al. 2019). In the following, we will briefly introduce these attribution methods, and then present their reformulations. We also discuss several popular attribution methods that cannot be unified into the framework. Note that the expansion point is the input sample $\boldsymbol{x}$ if not specified. All the proofs of theorems are in the Supplementary materials [2].

Gradient*Input. The attribution in Gradient*Input is calculated by multiplying the partial derivatives (of the output to the input) with the input, i.e., $a_i = f_{x_i}(\boldsymbol{x}) x_i$.

Theorem 1. *Gradient*Input can be reformulated as a first-order Taylor attribution w.r.t the baseline point $\tilde{\boldsymbol{x}} = \mathbf{0}$,*

$$a_i = T_i^{\alpha}.$$

That is, the attribution of x_i in Gradient*Input is $f_{x_i}\Delta_i$.

Occlusion-1. Occlusion-1 attribution calculates how much the prediction changes induced by occluding feature x_i with a zero baseline. The new occluded input is written as $\boldsymbol{x}|_{x_i=0}$. Then the attribution of feature x_i is defined as the change of the output, $a_i = f(\boldsymbol{x}) - f(\boldsymbol{x}|_{x_i=0})$.

Theorem 2. *The attribution of x_i in Occlusion-1 can be reformulated as the sum of first-order and high-order independent terms of x_i at baseline point $\tilde{\boldsymbol{x}} = \boldsymbol{x}|_{x_i=0}$,*

$$a_i = T_i^{\alpha} + T_i^{\gamma_d}.$$

The attribution of x_i in Occlusion-1 is $f_{x_i}\Delta_i + \frac{1}{2} f_{x_i x_i}\Delta_i^2$ in the second-order Taylor attribution.

Occlusion-patch. The attribution in Occlusion-patch (Occlusion-p for short) is conducted on a patch level. It constructs a zero patch baseline $\boldsymbol{x}|_{\boldsymbol{p_i}=\mathbf{0}}$ by occluding an image patch $\boldsymbol{p_j}$, and defines the prediction change $f(\boldsymbol{x}) - f(\boldsymbol{x}|_{\boldsymbol{p_j}=\mathbf{0}})$ as the attribution of feature in $\boldsymbol{p_j}$.

Theorem 3. *The attribution of $x_i \in \boldsymbol{p_j}$ in Occlusion-p can be reformulated as the sum of first-order, high-order independent terms of features in patch $\boldsymbol{p_j}$, and all high-order interactive terms involving the features in patch $\boldsymbol{p_j}$,*

$$\begin{aligned} a_i &= T_{\boldsymbol{p_j}}^{\alpha} + T_{\boldsymbol{p_j}}^{\gamma_d} + T_{\boldsymbol{p_j}}^{\gamma_t} \\ &= \sum\nolimits_{x_i \in \boldsymbol{p_j}} T_i^{\alpha} + \sum\nolimits_{x_i \in \boldsymbol{p_j}} T_i^{\gamma_d} + \sum\nolimits_{A \subset \boldsymbol{p_j}} T_A^{\gamma_t}. \end{aligned}$$

Particularly, the a_i is $\sum_{i \in \boldsymbol{p_j}} f_{x_i}\Delta_i + \sum_{i \in \boldsymbol{p_j}} \frac{1}{2} f_{x_i x_i}\Delta_i^2 + \sum_{i \in \boldsymbol{p_j}} \sum_{j \in \boldsymbol{p_j}} f_{x_i x_j}\Delta_i\Delta_j$ in second-order setting.

DeepLIFT and ϵ-LRP. DeepLIFT and ϵ-LRP compute relevance scores by using a recursive relevance propagation in a layer-wise manner. In DeepLIFT Rescale rule, $x_i^{(l)}$ and $x_j^{(l+1)}$ denote the neuron i at l-th layer and the neuron j at $(l+1)$-th layer, respectively, and $x_j^{(l+1)} = \sigma(\sum_i w_{ji} x_i^{(l)} + b_j)$. Here w_{ji} is the weight parameter, b_j is the additive bias, and σ is a non-linear activation function. DeepLIFT propagates the output difference between an input $\boldsymbol{x}$ and a baseline $\tilde{\boldsymbol{x}}$ to the input layer, and it calculates the relevance score of $x_i^{(l)}$ to $x_j^{(l+1)}$, denoted as $a_{ij}^{(l)}$, by

$$a_{ij}^{(l)} = \frac{z_{ji}^{(l)} - \tilde{z}_{ji}^{(l)}}{\sum_{i'} z_{ji'}^{(l)} - \sum_{i'} \tilde{z}_{ji'}^{(l)}} a_j^{(l+1)},$$

where $z_{ji}^{(l)} = w_{ji} x_i^{(l)}$ is the weighted impact of $x_i^{(l)}$ to $x_j^{(l+1)}$, analogously $\tilde{z}_{ji}^{(l)} = w_{ji} \tilde{x}_i^{(l)}$ denotes the weighted impact of the baseline, and $a_j^{(l+1)} = \sum_k a_{jk}^{(l+1)}$ denotes the total relevance score of neuron $x_j^{(l+1)}$ to all neurons in $(l+1)$-th layer. The formula of ϵ-LRP is similar to DeepLIFT, please see details in the supplementary materials.

Theorem 4. *The relevance score of $x_i^{(l)}$ to $x_j^{(l+1)}$ in DeepLIFT can be reformulated as the weighted sum of first-order term of $x_i^{(l)}$ and all high-order terms at baseline $\tilde{x}_i^{(l)}$,*

$$a_{ij}^{(l)} = T_i^{\alpha} + \frac{z_{ji}^{(l)} - \bar{z}_{ji}^{(l)}}{\sum_{i'} z_{ji}^{(l)} - \sum_{i'} \bar{z}_{ji}^{(l)}} T^{\gamma}.$$

In the second-order setting, the attribution $a_{ij}^{(l)}$ is $f_{x_i}\Delta_i + \frac{z_{ji}^{(l)} - \bar{z}_{ji}^{(l)}}{\sum_{i'} z_{ji}^{(l)} - \sum_{i'} \bar{z}_{ji}^{(l)}} (\sum_i \frac{1}{2} f_{x_i x_i}\Delta_i^2 + \sum_{ij} f_{x_i x_j}\Delta_i\Delta_j)$.

[2] The supplementary materials could be downloaded at: https://arxiv.org/abs/2008.09695

Integrated Gradients. The attribution in Integrated Gradients integrates the gradients along the straight line path from a baseline point $\tilde{\boldsymbol{x}}$ to an input $\boldsymbol{x}$. The points along the path are denoted as $\boldsymbol{x'} = \tilde{\boldsymbol{x}} + \alpha(\boldsymbol{x} - \tilde{\boldsymbol{x}}), \alpha \in [0, 1]$. The attribution of feature x_i is computed by

$$a_i = (x_i - \tilde{x}_i) \int_0^1 \frac{\partial f(\tilde{\boldsymbol{x}} + \alpha(\boldsymbol{x} - \tilde{\boldsymbol{x}}))}{\partial x_i} d\alpha. \quad (1)$$

Theorem 5. *The attribution of x_i in Integrated Gradients can be reformulated as the sum of first-order term of x_i, high-order independent terms of x_i, and an assignment from high-order interactive terms involving x_i at baseline $\tilde{\boldsymbol{x}}$,*

$$a_i = T_i^{\alpha} + T_i^{\gamma_d} + a_i^{\gamma_t},$$

where $a_i^{\gamma_t} = \sum_{K=2}^{\infty} \sum_{k_i} \frac{k_i}{K} (\Delta_i^{k_i} \prod_{\sum_j k_j = K - k_i} C \Delta_j^{k_j})$ is the assignment, and $C = \frac{1}{K!} \binom{K}{k_1, \ldots, k_n} \frac{\partial f(\boldsymbol{x})}{\partial x_1^{k_1} \ldots \partial x_i^{k_i} \ldots \partial x_n^{k_n}}$ is the Taylor expansion coefficient of $\Delta_1^{k_1} \ldots \Delta_i^{k_i} \ldots \Delta_n^{k_n}$.

In brief, Integrated Gradients allocates $\frac{k_i}{K}$ proportion of the high-order interactive term $\Delta_1^{k_1} \ldots \Delta_i^{k_i} \ldots \Delta_n^{k_n}$ to x_i. In the second-order setting, the attribution of x_i in Integrated Gradients is $f_{x_i} \Delta_i + \frac{1}{2} f_{x_i x_i} \Delta_i^2 + \frac{1}{2} \sum_{j \neq i} f_{x_i x_j} \Delta_i \Delta_j$.

Expected Gradients. Expected Gradients is an extension of Integrated Gradients. Expected Gradients samples baseline points from a prior distribution $p_D(\tilde{\boldsymbol{x}})$, instead of specifying only one baseline point in Integrated Gradients. The attribution a_i is then computed by integrating the Integrated gradients attributions along the baseline distribution,

$$a_i = \int_{\tilde{\boldsymbol{x}}} p_D(\tilde{\boldsymbol{x}})(x_i - \tilde{x}_i) \int_0^1 \frac{\partial f(\tilde{\boldsymbol{x}} + \alpha(\boldsymbol{x} - \tilde{\boldsymbol{x}}))}{\partial x_i} d\alpha d\tilde{\boldsymbol{x}}. \quad (2)$$

A common choice for prior distribution is to add a zero-mean, independent Gaussian distribution to $\boldsymbol{x}$. That is, $\tilde{\boldsymbol{x}} \sim N(\boldsymbol{x}, \sigma^2)$.

We can see the attribution in Expected Gradients, Eq. 2, is an integral over the baselines distribution of the attribution in Integrated Gradients, Eq. 1. Therefore, we can have a direct corollary that the attribution in Expected Gradients can also be reformulated into the Taylor attribution framework, and its reformulation is a integral of the reformulation of Integrated Gradients in theorem 5. Please see the specific reformulation and derivation in the supplementary materials.

In additional to the seven attribution methods, some attribution methods have been proven that they are the first-order Taylor attributions at (well-chosen) nearest root point, such as LRP-$\alpha\beta$ (Montavon et al. 2019) and DeepTaylor (Montavon et al. 2017). In the mean time, there are also several popular attribution methods that cannot be unified into the Taylor attribution framework, e.g., Deconvnet (Zeiler and Fergus 2014) and Guided BP (Springenberg et al. 2014). The two attribution methods do not analyze the output change $\Delta f = f(\boldsymbol{x}) - f(\tilde{\boldsymbol{x}})$ as the proposed framework does. It has been theoretically and empirically shown they are to recover the input (Adebayo et al. 2018; Nie, Zhang, and Patel 2018).

Theoretical Analysis of the Attribution Methods

The proposed Taylor reformulations enable us to examine rationales, measure fidelity, and analyze limitation for the attribution methods in a systematic and theoretical way. Firstly, we find **Gradient*Input**, **Occlusion-1**, and **Occlusion-patch** have a large approximation error ϵ (i.e., low fidelity) as they all fail to completely reflect the high-order Taylor terms. Theorem 1 shows Gradient*Input is a first-order Taylor attribution, which only takes the first-order terms into consideration. Although Occlusion-1 partially characterizes the high-order independent effects in Theorem 2, it fails to attribute the interactive terms. The complex interactions among features (pixels) always contains critical information for prediction in DNNs. Theorem 3 shows Occlusion-patch considers the overall effects of the features in the patch, including both independent and interactive terms. However, it assigns the same contribution score to all features in the patch, which fails to provide fine-grained attributions. Moreover, the interactive effects among different patches are neglected in Occlusion-patch.

From the reformulations, we can see **DeepLIFT**, **ϵ-LRP**, and **Integrated Gradients** have a small approximation error ϵ as they attribute the high-order terms. DeepLIFT and ϵ-LRP assigns weighted high-order terms of all features to feature x_i, which obviously fails to distinguish the high-order contributions of different features. As DeepLIFT and ϵ-LRP conduct in a layer-wise manner, the impact of high-order terms in each layer is much smaller than the one in a network, which may relieve the incorrect assignment problem.

Integrated Gradients is an average of first-order derivatives along the path. However, Theorem 5 shows that Integrated Gradients not only attributes the first and high-order independent terms, but also correctly assigns the interactive terms. Hence it's considered that Integrated Gradients is a superior attribution. This theoretical finding may provide an insight into why Integrated Gradients can well identify important features in input image. The performance of Integrated Gradients highly depends on baseline. However, it is difficult to select a good baseline. If use a black image as a baseline, integrated gradients will not highlight black pixels as important even if they make up the object of interest.

Lastly, we discuss about **Expected Gradients**. Expected Gradients relieved the problem induced by baseline by averaging among multiple baselines sampled from a Gaussian distribution, as analyzed in section 2.4. It reduces the probability that the attribution is dominated by a specific baseline.

Three Principles of Attribution

Based on the reformulations and analysis, we find a good Taylor attribution depends on three key factors: i) the Taylor approximation error $\epsilon(\tilde{\boldsymbol{x}}, K)$; ii) whether the Taylor terms in $g_K(\boldsymbol{x}, \Delta)$ are assigned correctly; iii) the baseline point $\tilde{\boldsymbol{x}}$. Accordingly, we establish three principles of a good attribution and advocate the principles should be followed by other attribution methods.

First principle: After Taylor reformulating, an attribution method should *has a low approximation error* $\epsilon(\tilde{\boldsymbol{x}}, K), \forall \tilde{\boldsymbol{x}}$. This principle is similar to the completeness axiom.

Principles	First	Second	Third
Gradient*Input			
Occlusion-1 & -p		partially	
DeepLIFT & ϵ-LRP	$\surd$		
Integrated	$\surd$	$\surd$	
Expected	$\surd$	$\surd$	$\surd$

Table 2: A summary of the principles followed by the attribution methods.

Second principle: After Taylor reformulating, an attribution method should *correctly assign the independent and interactive terms*. For instance, the first-order and high-order independent terms of feature x_i should only be assigned to a_i, and the interactive terms of features in subset A should only be attributed to the features in subset A.
Third principle: After Taylor reformulating, an attribution method should choose *an unbiased baseline*.

Table 2 presents a summary of the principles followed by the attribution methods, in which Occlusion-1 and Occlusion-p partially follow the second principle because they partially attribute high-order terms, as discussed above.

Experiments

In this section, we empirically validate our Taylor reformulations, and then investigate the relationship between attribution performance and the three principles via benchmarking on MNIST and Imagenet. We use GI, Occ-1, Occ-p, DL, IG, and EG to denote Gradient*Input, Occlusion-1, Occlusion-p, DeepLIFT Rescale, Integrated Gradients, and Expected Gradients, respectively.

Empirical Validations of Reformulations

In this section, we empirically validate our Taylor reformulations by comparing the attribution results from the original attribution methods and the corresponding Taylor reformulations. We firstly compute the attribution vector by the original attribution method and their Taylor reformulation. Then we adopt average percentage change as the metric to measure the difference between the attributions,

$$d = \frac{1}{N}\sum_{i=1}^{N}\frac{||\boldsymbol{a}_o(i)-\boldsymbol{a}_r(i)||^2}{||\boldsymbol{a}_o(i)||^2}\times 100\%,$$

where $\boldsymbol{a}_o(i)$ and $\boldsymbol{a}_r(i)$ denote the attribution vectors obtained by the original attribution method and the corresponding Taylor reformulation, respectively, for sample i. And N is the number of samples in the dataset. As DeepLIFT attributes in a layer-wise manner, we compute the percentage change of each layer and then average them.

We conduct the validation experiments on three models: i) **Poly**, a second-order polynomial model. ii) **M-sg**, a three-layer multi-layer perceptron (MLP) model with sigmoid activation, iii) **C-sg**, a three-layer CNN model with sigmoid activation. These models are all trained on MNIST [3] dataset.

[3] http://yann. lecun. com/exdb/mnist/

Models	Poly	M-sg	C-sg
GI	0	0	0
Occ-1	0	0.05%	13.17%
Occ-2×2	0	0.35%	21.22%
Occ-4×4	0	2.00%	35.93%
DL	0	2.35%	32.76%
IG	0.12%	3.50%	40.29%

Table 3: Average percentage changes between original attribution methods and their Taylor reformulations.

The average percentage change metrics are averaged on 3k validation set. We use second-order Taylor reformulations to validate the theories as higher-order one is always computationally intractable. Our theoretical results would expect the average percentage change metric should be small for the models which could be well approximated by second-order Taylor expansion.

Table 3 lists the average percentage change metrics of the aforementioned attribution methods. Occlusion-p is implemented for patch size 2×2 and 4×4. It can be seen that the metrics between attribution methods and their Taylor reformulations are equal to 0 in second-order Poly model (The marginal difference of Integrated Gradients is due to the integral approximation error) and M-sg models. This demonstrates the correctness of our theoretical reformulations.

The metrics on C-sg model are obviously larger than other models. We find that the discrepancy is mainly due to the incapability of second-order Taylor expansion to approximate C-sg model, instead of the proposed reformulations. We compute that the average (normalized) approximation error $\bar{\epsilon} = \frac{1}{N}\sum_i |\epsilon_i|/|\Delta f_i|$ of second-order Taylor expansion to Poly, M-sg, and C-sg model are 0, 0.084, 0.420 respectively. We observe that C-sg model has the largest approximation error. Moreover, there is a strong correlation between the percentage change and approximation error, which implies that the percentage change is resulted by the approximation error. For example, the correlation coefficient for IG method on C-sg model is 0.72.

Attribution Assessment

To investigate the relationship between attribution performance and three principles, we benchmark the six attribution methods in terms of infidelity and object localization accuracy on MNIST and Imagenet (Russakovsky et al. 2015) [4].

We evaluate the infidelity of these attribution methods on images from MNIST dataset. We adopt the infidelity metric proposed in (Yeh et al. 2019), which quantifies the degree to which it captures how the predictor function itself changes in response to significant perturbations. We use the square removal perturbation to assess the attributions of MLP, C-sg,

[4] ϵ-LRP has been shown equivalent to GI theoretically, so we do not show the results of ϵ-LRP.

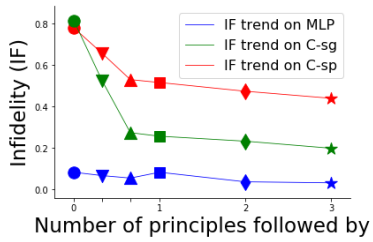

Figure 2: Infidelity trends on MLP (blue), C-sg (green), and C-sp models (red). The circle, triangle down, triangle up, square, diamond, and star denotes GI, Occ-1, Occ-p, DL, IG, and EG, respectively. x-axis denotes the number of principles the methods followed by, and y-axis denotes infidelity.

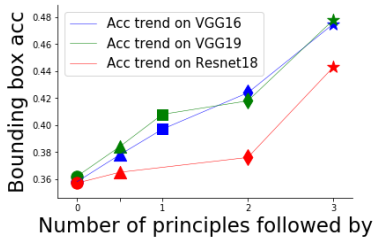

Figure 3: Bounding box accuracy trends on VGG16 (blue), VGG19 (green), and Resnet18 models (red). x-axis denotes the number of principles followed by, and y-axis denotes Bounding box accuracy.

and C-sp (CNN softplus) models. The infidelity trends (Figure 2) show a negative correlation between the infidelity and the number of principles followed by these methods. Furthermore, we also compare the infidelity of first and second-order Taylor explanations on MNIST dataset. Experimental results show that the average infidelities of the second-order one (0.27 on C-sg, 0.75 on C-sp) are significantly lower than the infidelity of the first-order one (0.48 on C-sg, 0.84 on C-sp). The results indicate that additional high-order Taylor terms indeed help improve the explanation.

We also investigate the relationship by measuring the attribution performance of these attribution methods on Imagenet. Here bounding box accuracy (Schulz et al. 2020) is adopted to evaluate how well attribution methods locate objects of interest. Assume the annotated bounding box contains n pixels. We select top n pixels according to ranked attribution scores and count the number of pixels m inside the bounding box. The ratio $\frac{m}{n}$ is used as the metric of localization accuracy. We only consider the images whose bounding boxes cover less than 33% of the input image. The bounding box accuracies are calculated on VGG16, VGG19, and Resnet18 networks. The trends in Figure 3 shows a positive correlation between the bounding box accuracy and the number of principles followed by these method.

The visualization comparisons among attribution methods are shown in Figure 4. Firstly, the saliency maps based on Gradient and GI are visually noisy and involves some irrelevant regions to the prediction. Occ-p, which incorporates high-order terms into the attribution, can approximately highlight the object of interest. DL and IG accurately identify right location of object, however they are very sensitive to the selection of baseline. For example, When interpreting corgi image with a black image baseline, IG assigns significantly higher contributions to white areas than black areas. This is due to the bias induced by large differences of input change Δ among different feature dimensions. EG solved this issue, and its generated saliency maps, evenly distributed with less noises, are sharper, and clearly display the shapes and boarders of the objects.

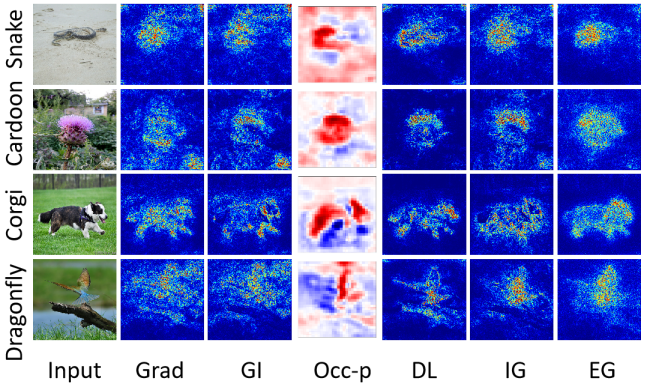

Figure 4: Visualization saliency map comparison.

Related Work

There are a few works on understanding the theoretical groundings of attribution methods. Deconvnet and Guided BP have been theoretically proved (Nie, Zhang, and Patel 2018) that they are essentially doing (partial) image recovery, which is unrelated to decision making. Some efforts have been devoted to unifying existing attribution methods recently. Several attribution methods are unified under the framework of additive feature attribution (Lundberg and Lee 2017), or reformulated as multiplying a modified gradient with input (Ancona et al. 2018) or summarized as first-order Taylor decomposition on different baseline points (Samek et al. 2020). To our knowledge, this is the first work to unify these attribution methods and further analyze the interactive Taylor terms by high-order Taylor decomposition.

Conclusion and Future Work

In this work, we propose a general Taylor attribution framework and theoretically reformulate several mainstream attribution methods into the attribution framework. Based on reformulations, we systematically analyze the attribution methods in terms of their rationale, fidelity, and limitation. Based on the reformulation and analysis, we establish three principles for a good attribution. In the future work, we will promote a more general Taylor framework, i.e., unify more attribution methods that analyze or decompose the output difference into the proposed framework, such as Shapley value, DeepLIFT RevealCancel, and so on.

Acknowledgements

The authors thank the anonymous reviewers for their helpful comments. This work is partially supported by the NSFC under grants Nos. 61673018, 61272338, 61703443 and Guangzhou Science and Technology Founding Committee under grant No.201804010255 and Guangdong Province Key Laboratory of Computer Science.

References

Adebayo, J.; Gilmer, J.; Muelly, M.; Goodfellow, I.; Hardt, M.; and Kim, B. 2018. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, 9505–9515.

Ancona, M.; Ceolini, E.; Öztireli, C.; and Gross, M. 2018. Towards better understanding of gradient-based attribution methods for Deep Neural Networks. In *International Conference on Learning Representations*.

Bach, S.; Binder, A.; Montavon, G.; Klauschen, F.; Müller, K.-R.; and Samek, W. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one* 10(7).

Du, M.; Liu, N.; and Hu, X. 2019. Techniques for interpretable machine learning. *Communications of the ACM* 63(1): 68–77.

Erion, G.; Janizek, J. D.; Sturmfels, P.; Lundberg, S.; and Lee, S.-I. 2019. Learning explainable models using attribution priors. *arXiv preprint arXiv:1906.10670* .

Lundberg, S. M.; and Lee, S.-I. 2017. A unified approach to interpreting model predictions. In *Advances in neural information processing systems*, 4765–4774.

Montavon, G.; Binder, A.; Lapuschkin, S.; Samek, W.; and Müller, K.-R. 2019. Layer-wise relevance propagation: an overview. In *Explainable AI: interpreting, explaining and visualizing deep learning*, 193–209. Springer.

Montavon, G.; Lapuschkin, S.; Binder, A.; Samek, W.; and Müller, K.-R. 2017. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition* 65: 211–222.

Nie, W.; Zhang, Y.; and Patel, A. 2018. A theoretical explanation for perplexing behaviors of backpropagation-based visualizations. In *Proceedings of the 35th International Conference on Machine Learning-Volume 70*, 3809–3818. JMLR. org.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A. C.; and Fei-Fei, L. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115(3): 211–252. doi:10.1007/s11263-015-0816-y.

Samek, W.; Montavon, G.; Lapuschkin, S.; Anders, C. J.; and Müller, K.-R. 2020. Toward Interpretable Machine Learning: Transparent Deep Neural Networks and Beyond. *arXiv preprint arXiv:2003.07631* .

Samek, W.; Montavon, G.; Vedaldi, A.; Hansen, L. K.; and Müller, K.-R. 2019. *Explainable AI: interpreting, explaining and visualizing deep learning*, volume 11700. Springer Nature.

Schulz, K.; Sixt, L.; Tombari, F.; and Landgraf, T. 2020. Restricting the Flow: Information Bottlenecks for Attribution. In *International Conference on Learning Representations*.

Shrikumar, A.; Greenside, P.; and Kundaje, A. 2017. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 3145–3153. JMLR. org.

Shrikumar, A.; Greenside, P.; Shcherbina, A.; and Kundaje, A. 2016. Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713* .

Smilkov, D.; Thorat, N.; Kim, B.; Viégas, F.; and Wattenberg, M. 2017. Smoothgrad: removing noise by adding noise. In *International Conference on Learning Representations Workshop*.

Springenberg, J. T.; Dosovitskiy, A.; Brox, T.; and Riedmiller, M. 2014. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806* .

Sturmfels, P.; Lundberg, S.; and Lee, S.-I. 2020. Visualizing the impact of feature attribution baselines. *Distill* 5(1): e22.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 3319–3328. JMLR. org.

Yeh, C.-K.; Hsieh, C.-Y.; Suggala, A.; Inouye, D. I.; and Ravikumar, P. K. 2019. On the (in) fidelity and sensitivity of explanations. In *Advances in Neural Information Processing Systems*, 10967–10978.

Zeiler, M. D.; and Fergus, R. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*, 818–833. Springer.

Zintgraf, L. M.; Cohen, T. S.; Adel, T.; and Welling, M. 2017. Visualizing deep neural network decisions: Prediction difference analysis. *arXiv preprint arXiv:1702.04595* .