

Supplementary Materials: Effective optimization of root selection towards improved explanation of deep classifiers

Anonymous Authors

1 ABSTRACT

This document accompanies the paper "Effective optimization of root selection towards improved explanation of deep classifiers". In this document, we detail the DTD (Deep Taylor Decomposition) details and provide supplementary experimental results. We share our codes via the link: <https://github.com/AnonymousSCI/LearnRootDTD>

2 DEEP TAYLOR DECOMPOSITION

This section provides the necessary theoretical foundation of DTD, including our analysis, to introduce the background and foundation for our proposed.

2.1 Relevance propagation

Recalling DTD, Let's assume $R_{j,c}^{l+1}$ represents the relevance of a neuron x_j^{l+1} with respect to class c . x_i^l is a previous layer neuron in the l -th layer and connected to x_j^{l+1} . The superscript l is the index of the layer in a neural network which will be omitted when it does not lead to ambiguity, and the subscript i denotes the index of the neuron within that layer.

If x_j is in the output layer, then $R_{j,c} = x_j$. Due to the confusion of class-related information in x_i , direct computation of $R_{i,c}$ is not feasible. Therefore, DTD assumed the existence of a mapping function between x_i and $R_{j,c}$, which can be approximated using a Taylor expansion:

$$R_{j,c} = \sum_i \underbrace{\frac{\partial R_{j,c}}{\partial x_i}}_{R_{i \leftarrow j,c}} |_{\tilde{x}_i^{(j)}} (x_i - \tilde{x}_i^{(j)}), \quad (1)$$

where $R_{i \leftarrow j,c}$ refers to the relevance that is distributed from x_j to x_i . To this end, $R_{i,c}$ can be expressed as the aggregation of all $R_{i \leftarrow j,c}$ connected to x_i :

$$R_{i,c} = \sum_j \frac{\partial R_{j,c}}{\partial x_i} |_{\tilde{x}_i^{(j)}} (x_i - \tilde{x}_i^{(j)}). \quad (2)$$

where $\tilde{x}_i^{(j)}$ is the root chosen for neuron x_j that makes the output $f(\tilde{x}_i)$ equal to zero. Montavon et al. [5] derived the layer-specific roots at the plane $f(\tilde{x}_i) = 0$. Specifically, taking the commonly used function: $x_j = \text{Relu}(w_j^T x_i + b_j)$ as an example, the root can be obtained from the intersection of:

$$\begin{cases} \tilde{x}_i = x_i + t v_j \\ w_j^T \tilde{x}_i + b_j = 0. \end{cases} \quad (3)$$

By calculating t , the root can be derived as $\tilde{x}^{(j)} = x_i - \frac{w_j^T x_i + b_j}{w_j^T v_j} v_j$, it can then be inserted into Equation (2) to obtain:

$$R_{i,c} = \sum_j w_j \odot \frac{w_j^T x + b_j}{w_j^T v_j} v_j = \sum_j w_j \odot \frac{v_j}{w_j^T v_j} R_{j,c}. \quad (4)$$

Depending on the inputs and activation functions, the direction of v_j can vary [6], leading to distinct propagation rules (e.g. DTD- ω^2). Specifically, when the root is equal to zero, the DTD simplifies to the product of the gradient and the input [6], which is equivalent to the LRP-0 rule [4].

The aforementioned computation process is performed on each neuron in every layer. As a result, relevance is propagated from the output layer all the way to the input layer, forming a heatmap that indicates the contribution of each pixel.

It is important to note that, in general, explanation methods based on relevance propagation are expected to satisfy the layer-wise relevance conservation [1, 5], which is presented as

$$\forall x : f(x) = \sum_{\text{pixel}} R_{\text{pixel},c} \quad (5)$$

and

$$\dots = \sum_j R_{j,c}^{l+1} = \sum_i R_{i,c}^l = \dots = \sum_{\text{pixel}} R_{\text{pixel},c}. \quad (6)$$

In the following, we demonstrate that the current DTD method partially fails to satisfy this conservation, leading to an inadequate representation of a neuron's contribution in terms of relevance.

2.2 Challenges of DTD

DTD is limited by the local linearization and root selection. In the following toy example, we will illustrate the limitations via a neural network that consists of the commonly used nonlinear function, i.e. Sigmoid. As shown in Figure 1, this neural network receives two pixels, x and y , as inputs.

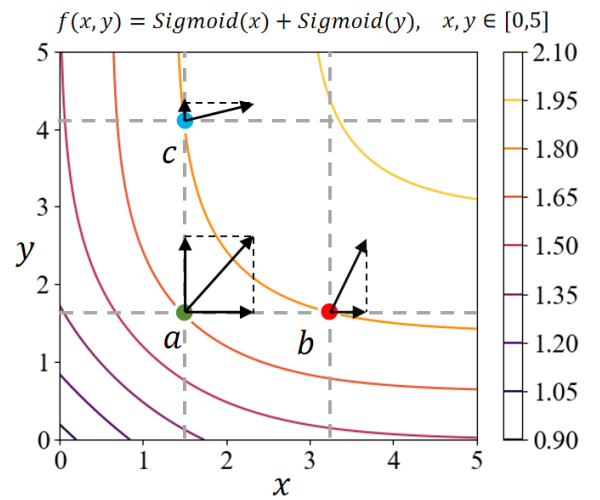


Figure 1: The output contour of the example network $f(x, y)$ which is mentioned in section 7.2.

Neighboring roots lose interpretability. When the root is in the neighborhood of x in Equation (2), $\Delta = x - \tilde{x}_i$ becomes a small constant, and the relevance is determined only by the nearby gradients. Given a point a within the domain of the neural network as defined in Figure 1, for all points to the right of a , the relevance of pixel y is should the same as that of a . Since $x_a < x_b$ and $\frac{\partial f}{\partial x_a} > \frac{\partial f}{\partial x_b}$, however, it can be proven that the existing of a point b such that:

$$\exists b : \Delta \times \frac{\partial f}{\partial x_a} > \Delta \times \frac{\partial f}{\partial x_b} \quad (7)$$

and

$$\exists b : x_a \times \frac{\partial f}{\partial x_a} = x_b \times \frac{\partial f}{\partial x_b}, \quad (8)$$

where $f(x_a, y_a) < f(x_b, y_b)$. Both Equation (7) (DTD within the neighborhood) and Equation (8) (LRP-0 rule) result in relevances that do not satisfy the conservation law as described in Equation (5). When the model detects an increase in object information in the pixel domain, the overall correlation tends to remain unchanged or even decreased.

The ideal root is being questioned. While all root derivations in DTD [4, 5] are based on Equation (3), the value of t is not constrained, resulting in many actual roots being far away from x_i [6]. In this scenario, the theory of Taylor expansion is not supported [3]. As illustrated in Figure 1, for example, the actual ideal roots are at an infinitely distant location in the lower left corner where the gradient is zero. Fortunately, LRP rules can be derived using Equation (3) and it is effective in most cases (Note in Equation (8), it is highly likely that the second b cannot be found in the right of a). However, the theoretical foundation of DTD is still being questioned for the following reasons: i) LRP rules is derived under the assumption that all layers are linear; ii) the different variants of LRP are inherently class-agnostic [2], contradicting the premise of Equation (1).

What is the root that minimizes the output of a function?

Here, we aim to demonstrate that limitations of DTD extend beyond the issue of roots and encompass the underfitting of nonlinear functions by first-order Taylor approximation. Returning to the toy example described in Figure 1, we manually specify an optimal root located at $(0, 0)$, where the model's output is minimized within the domain. As the root $(0, 0)$ is not far, let's assume that point c is located directly above a , while b is to the right of a , with its height (i.e. the model's output) equal to that of c . We can then establish that the difference in relevance value between a and b depends on the x -value, and the difference in relevance between a and c is dependent on the y -value. Through observations, it can be inferred that, with a being the reference, the relevance of point c is greater than that of b , even though they are actually on the same contour line. This observation contradicts Equation (5).

3 QUALITATIVE EVALUATION RESULTS

More visualizations of heatmaps are presented here.

Figure 2 illustrates the comparison of heatmaps generated by the proposed and all baselines, involving multiple classes of ImageNet. In these examples, the proposed method produces a less noisy interpretation, indicates that the proposed method correctly assigns a higher relevance to the object being detected.

We further compared the heatmaps generated by the proposed method with the suboptimal baseline (DTD- z^+) across four scenarios: feature-level interpretation (Figure 3), detail recovery (Figure 4), complex backgrounds (Figure 5), and multi-target scenes (Figure 6).

As can be seen from Figure 3, the heatmaps of proposed method significantly distinguishes the varying relevance among different regions of the detected object. This insightful heatmaps reveal the critical features of the target object that play a dominant role in classification. Through these heatmaps, we gain valuable insights into the model's discernment of feature importance and its impact on the classification outcomes.

Figure 5 illustrates several typical complex background scenarios that involves significant line interference or occlusion, resulting in a lot of noise in the heatmaps of the suboptimal baseline (DTD- z^+).

REFERENCES

- [1] Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. 2016. Layer-wise relevance propagation for neural networks with local renormalization layers. In *Artificial Neural Networks and Machine Learning-ICANN 2016: 25th International Conference on Artificial Neural Networks, Barcelona, Spain, September 6-9, 2016, Proceedings, Part II 25*. Springer, 63–71.
- [2] Hila Chefer, Shir Gur, and Lior Wolf. 2021. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 782–791.
- [3] Gary SW Goh, Sebastian Lapuschkin, Leander Weber, Wojciech Samek, and Alexander Binder. 2021. Understanding integrated gradients with smoothtaylor for deep neural network attribution. In *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 4949–4956.
- [4] Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. 2019. Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning* (2019), 193–209.
- [5] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. 2017. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern recognition* 65 (2017), 211–222.
- [6] Leon Sixt and Tim Landgraf. 2022. A Rigorous Study Of The Deep Taylor Decomposition. *Transactions on Machine Learning Research* (2022).

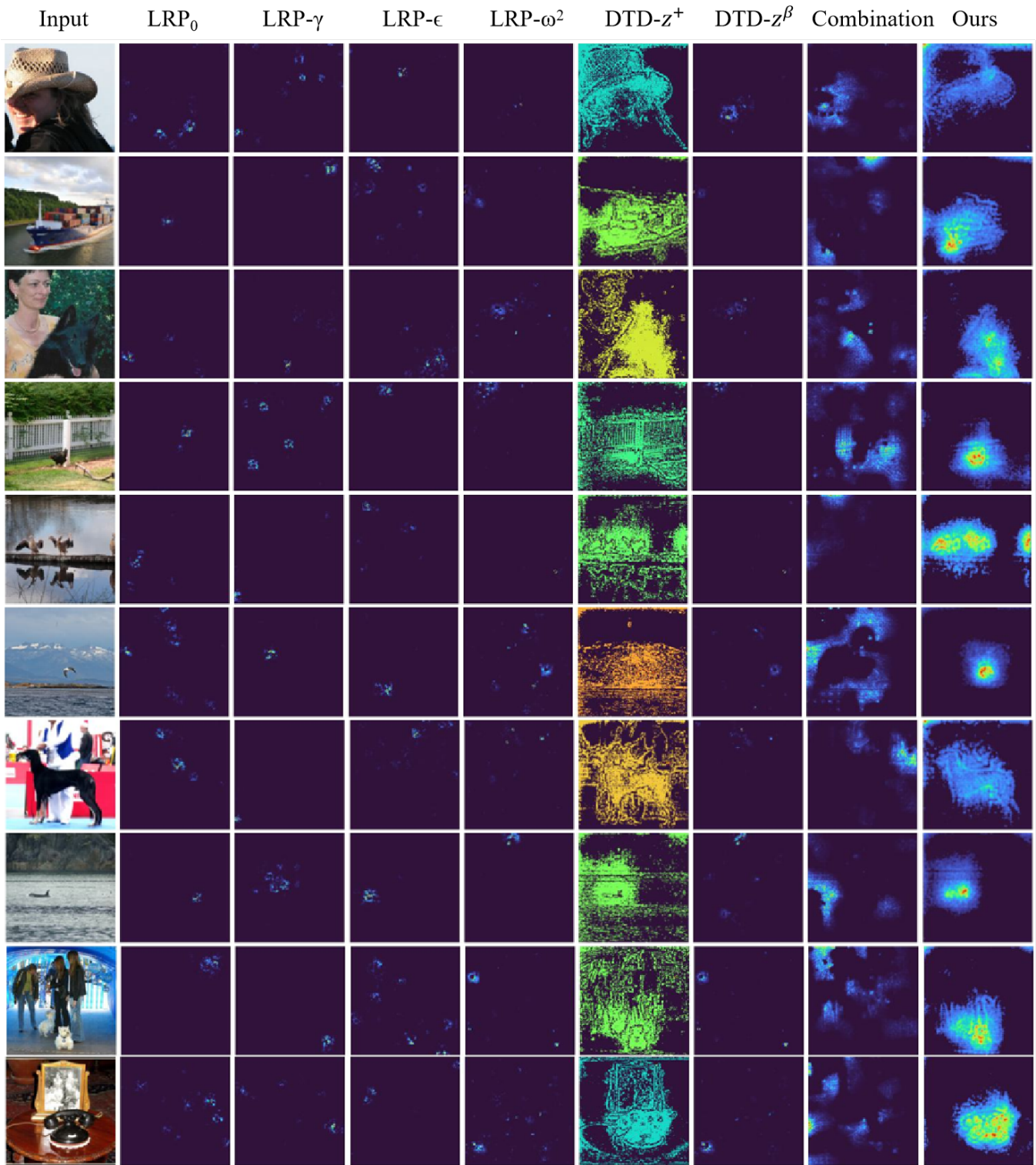


Figure 2: Illustration of the heatmap comparison between the proposed method and the baselines.

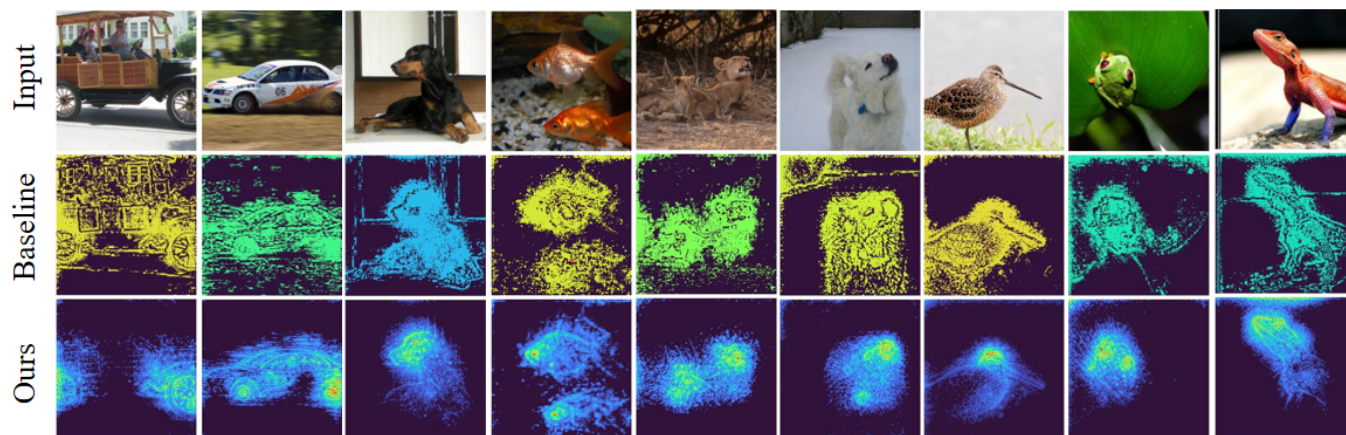


Figure 3: Illustration of the feature-level relevances that indicating the preference of ResNet18 for critical features.



Figure 4: Comparison of detail restoration ability between the suboptimal baseline and the proposed.

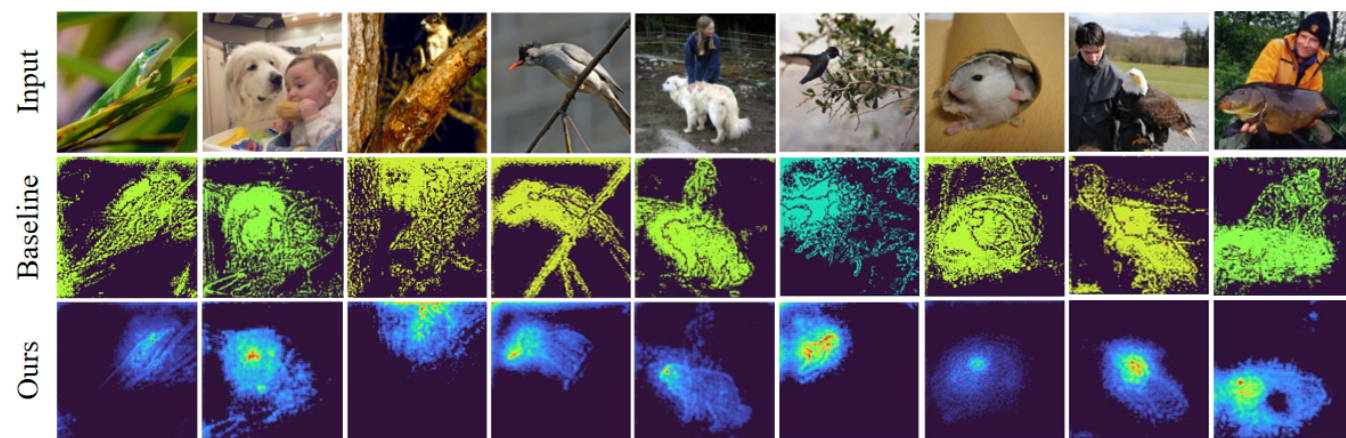


Figure 5: Comparison of heatmaps in complex background scenes.

