SmoothGrad Has Better Kendall's Rank for Removed Pixels

Bumjin Park

Abstract

Evaluating the input attributions of Deep Neural Network (DNN) is crucial for explanation of a black-box model. Recent work shows that the Integrated Gradients shows positive correlation for Kendall's rank correlation and cosine similarity with the perturbed input. However, input perturbation is a unclear baseline for the deviation of input. For example, SmoothGrad has already considers the input perturbation in its computation. To properly measure the input deviation, we propose masking based input deviation. We show that SmoothGrad and LRP has positive correlation with masked input, while IG fails to construct the strictly positive correlation.

1. Introduction

Recently, (Wang & Kong, 2022) shows that IG has positive relationship between Kendall's rank correlation and cosine similarity with the perturbed input data. However, it is unclear whether perturbation based deviation is a good choice. Even though the largely perturbed input has larger distance with the original input, the semantic distance could be lower when the perturbed pixel has small importance. For the better deviation of the input, we propose input importance based deviation for the similarity measure. We show that perturbation based input deviation has positive correlation for all input attribution methods, while the importance based perturbation has clear difference for SmoothGrad and IG.

2. Method

2.1. Kendall's Rank Correlation

Kendall's Rank Correlation measures rank correlation of two ordered sequences. Given $\{(x_1, y_1), \cdots (x_n, y_n)\}$, two pairs (x_i, y_i) and (x_j, y_j) are *concordant* if the order (x_i, x_j) and (y_i, y_j) agrees and discordant if the order disagrees. The Kendall's τ is defined by

$$\tau = \frac{(\#concordant) - (\#discordant)}{(\#pairs)} \tag{1}$$

$$= \frac{2}{n(n-1)} \sum_{i<j} \text{sign}(x_i - x_j)\text{sign}(y_i - y_j) \tag{2}$$

When two sequences perfectly agree, $\tau = 1$. In the case of perfect disagreement, $\tau = -1$ and independent sequences have $\tau = 0$. Kendall's rank correlation is particularly useful for sequences where the magnitude is less important and the order matters. Consider the input attribution of deep neural network such as Integrated Gradient, SmoothGrad, and LRP. These attributions may have different meaning for the quantity. However, two pixels have clear meaning; larger value has higher importance. Therefore, Rank correlation is good for correlation meausre of input attributions.

2.2. Attribution Methods

Integrated gradients (IG) (Sundararajan et al., 2017) computes the attribution with line integral of gradients from a baseline $\mathbf{a}$ to image $\mathbf{x}$

$$g(x)_i = (x_i - a_i) \times \int_0^1 \frac{f_y(\mathbf{a} + \alpha(\mathbf{x} - \mathbf{a}))}{\partial x_i} d\alpha \tag{3}$$

IG satisfies the axiom of completeness which guarantees $\sum_i g(\mathbf{x})_i = f_y(\mathbf{x}) - f_y(\mathbf{a}))$. As ususal, we use zero as the baseline, *i.e.*, $\mathbf{a} = \mathbf{0}$..

SmoothGrad (Smilkov et al., 2017) computes the sensitivity with the noised inputs

$$\hat{M}_c(x) = \frac{1}{n} \sum_1^n M_c(x + \mathcal{N}(0, \sigma^2)) \tag{4}$$

where $M_c(x) = \partial f_y(x)/\partial x$ is sensitivity. It is known that $10\% - 20\%$ noise balance the sharpness of sensitivity map and maintain the structure of the original image. Unlike IG, SmoothGrad uses the noise to the input when computing the input attribution.

Layer-wise Relevance Propagation (LRP) measures the contribution of individual pixels. The relavance score of j-th neuron in layer l from the layer $l+1$ is computed by

$$R_j^{(l)} = \sum_k \frac{w_{jk} a_j}{\sum_i w_{ik} a_i} R_k^{(l+1)} \tag{5}$$

where $R_j^{(l)}$ represents the relevance score of $j - th$ neuron of layer l with weight w_{ij}, and activation a_j. LRP holds the conservative law in the propagation.

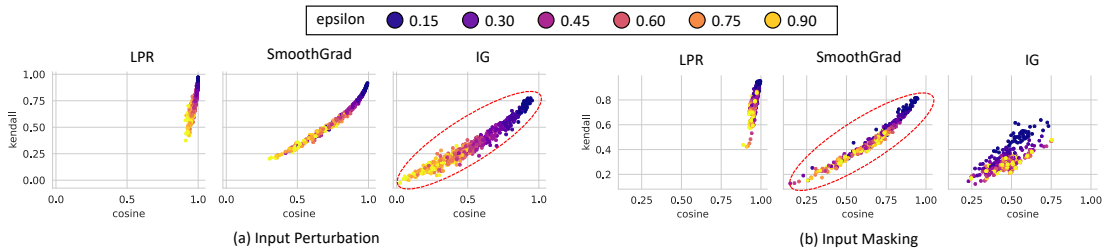

Figure 1. (a) Correlation and similarity for perturbed input $\hat{\mathbf{x}} = \mathbf{x} + \epsilon$ and original input x. IG shows clear correlation. (b) Correlation and similarity for perturbed input $\hat{\mathbf{x}} = M_{\mathbf{x}} \odot \mathbf{x}$ and original input x. SmoothGrad shows clear correlation.

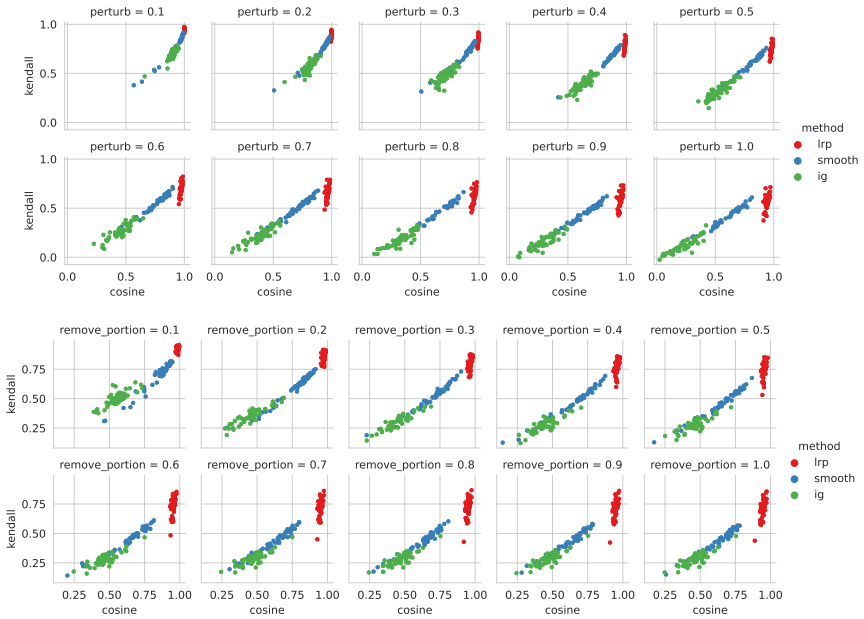

Figure 2. Scatter plot for each ϵ

2.3. Experiments

In the original experiment, the original input $\mathbf{x}$ is compared with the perturbed input $\hat{\mathbf{x}} = \mathbf{x} + \epsilon$. We reproduced result for 50 samples in CIFAR 10 dataset with trained 3 blocks CNN model. Only IG shows almost $\tau = 1$. However, the similarity with the input should consider the deviation from the input. One possible deviation is pixel removed image

$$\mathbf{x} = M_{\mathbf{x}} \odot \mathbf{x} \tag{6}$$

where $M_{\mathbf{x}}$ is the masking of the important regions obtained by the attribution map of the original input $\mathbf{x}$. Figure 1 shows the comparison of methods for different deviation methods of inputs. Pannel (a) shows that the input perturbation deviation shows positive correlation for all input attribution methods. IG shows the most positive correlation. However, it shows that the perturbation of the input has higher effects for IG while LRP and SmoothGrad preserves the attribution even though the input is slightly perturbed. Pannel (b) shows that IG fails to construct the linear correlation for removed portion ratio. That is, the IG do not preserves the linear relationship for the input deviation. SmoothGrad has $\tau = 1$ while the IG fails to show the less correlation for cosine similarity and Kendall's rank. Figure 2 shows distributions for each ϵ. When the masking ratio is smaller than 0.5, there is huge difference, while the masking ratio with larger than 0.6 results in similar results as the important samples are mostly removed.

3. Conclusion

Evaluating the input attribution methods should be done with careful understanding of the explanation bias. In this work, we show that masked inputs have positive correlation for the cosine similarity and the Kendall's rank correlation which demonstrate that simple perturbation of input is not enough deviation of the original input.

References

Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

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

Wang, F. and Kong, A. W.-K. Exploiting the relationship between kendall's rank correlation and cosine similarity for attribution protection. *arXiv preprint arXiv:2205.07279*, 2022.