Layer-wise Relevance Propagation for Neural Networks with Local Renormalization Layers

Alexander Binder, Gregoire Montavon, Sebastian Bach,
Klaus-Robert Muller, and Wojciech Samek

CRITIQUE

Presented by: Kurtis David, Harrison Keane, and Jun Min Noh

#1: Doesn't explore why Taylor is worse in some cases

dataset	methods	$\epsilon = 1$	$\epsilon = 0.01$	$\epsilon = 100$	$\beta = 1$	$\beta = 0$
Imagenet	$\mathrm{AUC}_{\mathrm{Taylor}} - \mathrm{AUC}_{\mathrm{identity}}$	-35.84	-26.84	8.47	0.29	1.98
MIT Places	$\mathrm{AUC}_{\mathrm{Taylor}} - \mathrm{AUC}_{\mathrm{identity}}$	-33.13	-24.59	5.34	-0.39	-1.06

Table 3. Impact of using the Taylor method in various settings. Negative value indicates that using the Taylor expansion for the local renormalization is better in AUC terms (i.e. heatmaps are more representative of the importance of each pixel).

For some values of ϵ and β, the Taylor method performs worse than identity rule for normalization layers. However, the authors do not further explain the reasoning behind the outcome.

#2: Figure 3

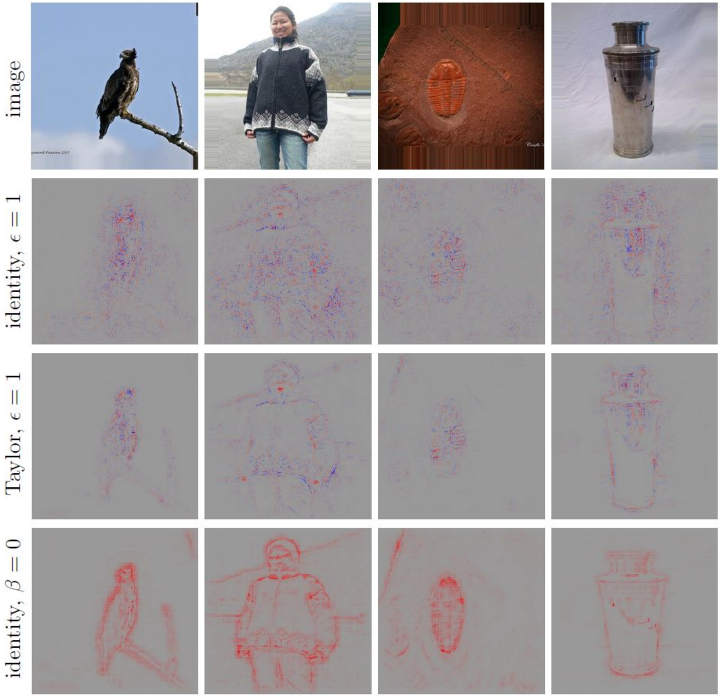

Fig. 3. Top row shows original unwarped image. Remaining rows show heatmaps produced by various parameters of the LRP method.

#3: Compresses too much with little gain

dataset	methods	$\Delta_{\epsilon=1}^{\epsilon=0.01}$	$\Delta_{\epsilon=0.01}^{\epsilon=100}$	$\Delta_{\epsilon=1}^{\beta=1}$	$\Delta_{\beta=1}^{\beta=0}$
Imagenet	identity	-21.29	2.75	-42.61	-49.07
	Taylor	-12.29	-41.75	-34.44	-50.76
MIT Places	identity	-20.19	12.91	-14.55	-49.37
	Taylor	-11.65	-22.55	-8.82	-48.7

Table 2. Comparison of different types of heatmap computations for Imagenet and MIT Places. We use the shortcut notation Δ_a^b for expressing $\mathrm{AUC_a} - \mathrm{AUC_b}$. Thus, a negative value indicates that the method produces better heatmaps with parameter a than with parameter b. Note that ϵ refers to equations 4 and 5; β refers to eq. 4 and 6.

#4: Not self-contained, lack of explanations

$$\Rightarrow y_k(z_1) \approx \frac{x_k}{(1+bx_k^2)^c} - 2bc \sum_{j:j\neq k} \frac{x_k x_j^2}{\left(1+b\sum_{i=1}^n x_i^2\right)^{c+1}} \tag{15}$$

$$y_k(x_1,\ldots,x_n) \approx \sum_{d=1}^{n} f_d(x)$$

$$y_k(x_1,\ldots,x_n) \approx \sum_{d=1}^{n} \frac{\partial y_k}{\partial x_d}(x^{(0)})(x_d - x_d^{(0)})$$

$$r_d(x) = R_k \cdot \frac{f_d(x)}{\sum_{d'=1}^{n} f_{d'}(x)}$$

#5: Need White Box Access

- For LRP, you need access to all weights