Improving Explainability of Integrated Gradients with Guided Non-Linearity

Hyuk Jin Kwon[1], Hyung Il Koo[2] and Nam Ik Cho[1]

hjkwon@ispl.snu.ac.kr, hikoo@ajou.ac.kr, nicho@snu.ac.kr

[1] Dept. of Electrical and Computer Engineering, Seoul National University
[2] Dept. of Electrical and Computer Engineering, Ajou University

Abstract In this paper, we present a new method that improves the measure of attribution and incorporates it into the integrated gradients method. To be precise, rather than using the conventional chain-rule, we propose a method called guided non-linearity that propagates gradients more effectively through non-linear units (e.g., ReLU and max-pool) so that only positive gradients backpropagate through non-linear units. Our method is inspired by the mechanism of action potential generation in postsynaptic neurons, where the firing of action potentials depends on the sum of excitatory (EPSP) and inhibitory postsynaptic potentials (IPSP). Experiments with 5 deep neural networks have shown that the proposed method outperforms others in terms of the deletion metrics and yields fine-grained and more human-interpretable attribution.

Motivation

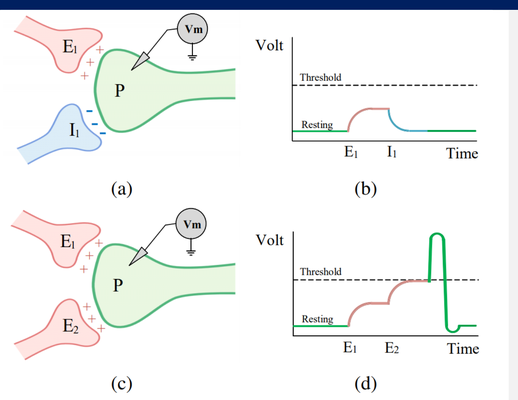

Fig 1. Diagram of synapses with different PSP

The conduction of action potential is controlled by voltages in the synaptic cleft. The excitatory post-synaptic potential, called EPSP, increase the postsynaptic potential and inhibitory post-synaptic potential, called IPSP, decrease the postsynaptic potential. We think that non-linear units (ReLU and max-pool) with positive gradients operate as EPSPs and the negative gradients as IPSPs.

In our postulations, we have to focus on positive gradients to find the cause of the current prediction.

Integrated Gradients

To make attribution of an input x for a given CNN (denoted as F), integrated gradients used a line integral of gradients along the path from a baseline image x' to the given input x :

$$\mathrm{IG}_i(x) = (x_i - x_i') \times \int_0^1 \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i} d\alpha$$

where i is used as a pixel index.

For the implementation, integrated gradients is approximated with its discrete version,

$$\mathrm{IG}_i(x) \approx \frac{1}{m} \sum_{k=1}^{m} \frac{\partial F(x' + \frac{k}{m} \times (x - x'))}{\partial x_i}.$$

Intuitively, IG is the sum of incremental contributions of the i-th pixel to the output (along the path from x' and x), where each contribution was evaluated with gradients.

Proposed Method

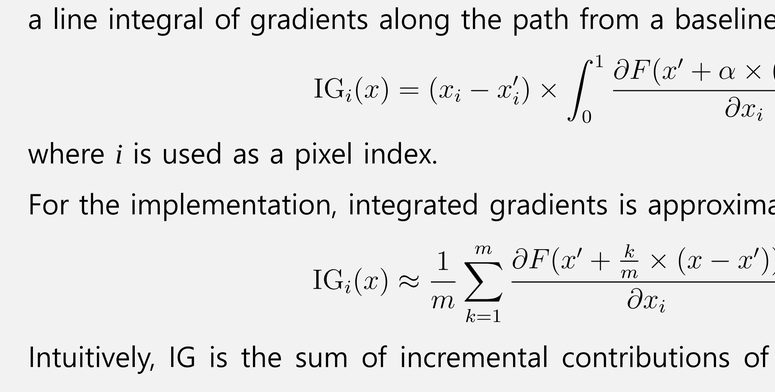

Fig 2. Comparison of our guided non-linearity with the normal backpropagation.

Based on our observation, it is natural to focus on the positive gradients in non-linear units (corresponding to axonal terminals) for attribution.

In other words, we have to focus on positive gradients to find the cause of the current prediction, since neurons yielding IPSPs are against the current prediction results.

We computationally achieve this goal by clipping negatively valued gradients in non-linear units to zero and use these new gradients in the path integral of IG,

For ReLU: $$\frac{\partial F(\cdot)}{\partial x} = \mathrm{relu}\left(\frac{\partial F(\cdot)}{\partial y} \odot I(x > 0)\right)$$

For max-pool: $$\frac{\partial F(\cdot)}{\partial x_{ij}} = \mathrm{relu}\left(\frac{\partial F(\cdot)}{\partial y_i} \odot I(x_{ij} = y_i)\right)$$

where $I(\cdot)$ is an indicator function, $\odot$ mean the elementwise product, i is the index for the output of max-pool and j is the index for the input.

Experimental Results

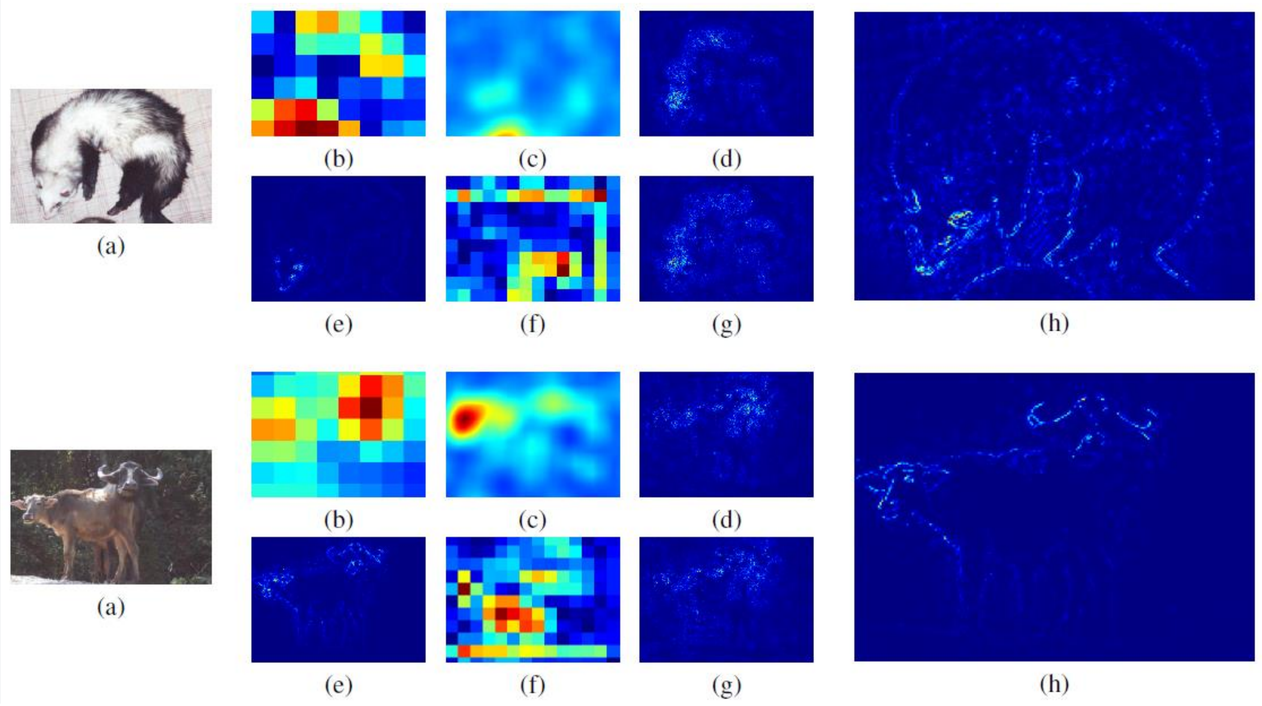

Fig 3. Comparison of attribution heatmaps using VGG16: (a) Input image (label: ferret, water buffalo), (b) Occlusion, (c) RISE, (d) Gradients, (e) Guided Backprop, (f) Grad-CAM, (g) Integrated Gradients (IG), and (h) ours.

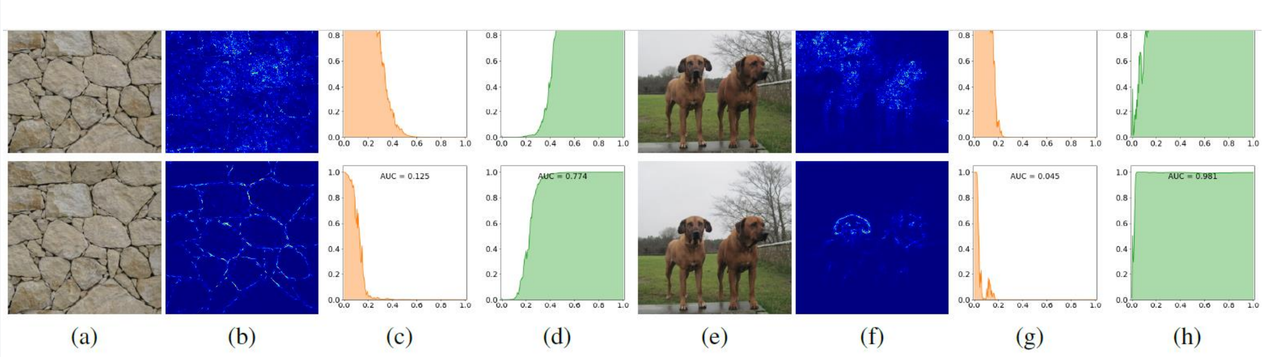

Fig 4. Illustrations of the deletion and insertion metrics for IG (upper) and our method (below) using ResNet50: (a), (e) input images (label: *stone wall*, *Rhodesian ridgeback*), (b), (f) attribution, (c), (g) curves for the deletion metric (IG: AUC=0.338/0.164, our method: AUC=0.125/0.045 respectively), (d), (h) curves for insertion metric (IG: AUC=0.597/0.774, our method: AUC=0.918/0.981 respectively).

	VGG16		VGG19		ResNet34		ResNet50		GoogleNet	
Methods	Deletion ↓	Insertion ↑	Deletion ↓	Insertion ↑	Deletion ↓	Insertion ↑	Deletion ↓	Insertion ↑	Deletion ↓	Insertion ↑
Occlusion [17]	0.1577	0.5755	0.1616	0.5770	0.1874	0.5914	0.2141	0.6309	0.1350	0.4667
LIME [28]	0.1014	0.6167	-	-	-	-	0.1217	**0.6940**	-	-
RISE [19]	0.0964	**0.6048**	0.0998	**0.6070**	0.1028	0.6308	0.1121	0.6762	0.0684	0.4995
Gradients [33]	0.0672	0.3270	0.0791	0.3423	0.1268	0.4221	0.1134	0.4234	0.0745	0.3574
GB [18]	0.0526	0.5279	0.0567	0.5445	0.0826	0.6141	0.0755	0.6460	0.0639	**0.5124**
GradCam [34]	0.1605	0.4305	0.1520	0.4578	0.1557	**0.6333**	0.1887	0.6715	0.1156	0.5086
IG [6]	0.0543	0.3621	0.0640	0.3792	0.1030	0.4575	0.0931	0.4589	0.0634	0.3936
Ours	**0.0495**	0.5151	**0.0532**	0.5295	**0.0763**	0.5932	**0.0721**	0.6295	**0.0601**	0.4912

Tab 1. Comparison of deletion and insertion metric for 5 networks.

We have evaluated the proposed method with 5 CNN architectures (VGG16 , VGG19, ResNet34, ResNet50 and GoogleNet) on 5,000 linearly sampled images from the validation split of ImageNet classification database. For the quantitative evaluation, we have used the deletion and insertion metrics which was designed to evaluate the quality of attribution without human intervention. As shown, our method outperforms IG in both metrics (Fig 4.) and outperforms all other methods in terms of the deletion metric and gets the insertion metric score comparable to the state-of-the-art results (Tab. 1). Although the proposed method yields a little lower insertion metric compared with perturbation methods, these methods lack the power to localize targets as can be seen in Fig 3-(b) and (c).

Conclusion

We have proposed an improved method that generates human-interpretable attribution. Our method modifies the back-propagation methods on ReLU and max-pool non-linearity used in the path integral of IG. Our method achieves the state-of-the-art deletion score and outperforms the IG method.