Backpropagation-based visualization methods for explaining CNNs

Christian Ickler

Friedrich Schiller University Jena, AI Summer School 2023

Mail: christian.ickler@uni-jena.de

ai.uni-jena.de

Introduction

In contemporary machine learning, the proficiency exhibited by artificial neural networks is indisputable; however, their decision-making processes are often difficult to comprehend. This lack of transparency poses significant problems, particularly in critical domains such as healthcare and law enforcement, where comprehending AI-driven decisions is of great importance to gain trust in the systems. This poster examines three popular backpropagation-based visualization methods for convolutional neural networks (CNNs): *DeepLIFT, Integrated Gradients* and *Grad-CAM*. These methodologies try to indicate pixel or feature relevance by using gradients of the output passed backwards [1]. After a detailed analysis of these techniques, a comparison is made to show their individual limitations. This aims to assist the reader in selecting a suitable method for specific use cases.

Methods

DeepLIFT

Deep Learning Important FeaTures (*DeepLIFT*) [2] assigns pixel-wise attribution scores based on the difference from an input image x to a reference image x'. The choice of reference should rely on domain-specific knowledge, aiming for a neutral prediction. Frequently, the black image or a blurred version of the input proves to be a suitable selection.

The contribution scores $C_{\Delta x_i \Delta t}$ of the difference-from-references of input neurons x_i to the difference-from-reference of the target output t satisfy the summation-to-delta property:

$$\sum_{i=1}^{n} C_{\Delta x_i \Delta t} = \Delta t$$

These contribution scores are averaged to obtain a multiplier that resembles a finite difference partial derivative:

$$m_{\Delta x \Delta t} = \frac{C_{\Delta x \Delta t}}{\Delta x}$$

The approach considers positive and negative contributions and can be calculated using one of the three proposed rules found in the *DeepLIFT* paper [2]:

- *Linear rule*: applicable to dense and convolutional layers without nonlinearities
- *Rescale rule*: applicable to single input nonlinear transformations, such as ReLU, sigmoid or tanh
- *RevealCancel rule*: alternative to the *Rescale rule*, serving as a fast approximation of Shapely values

The *DeepLIFT* paper proves the effectiveness of the chain rule for multipliers, enabling the computation of contribution scores for any neuron with respect to every target neuron within a single backward pass.

Integrated Gradients

Integrated Gradients [3], similar to *DeepLIFT*, assigns attribution scores to input features based on a neural network's predictions. It was designed using an axiomatic approach to fulfill two fundamental axioms for attribution methods:

- *Sensitivity*: if an input x and a baseline x', that differ in one feature, have different predictions, the attribution for this feature should not be zero
- *Implementation Invariance*: attribution consistency is maintained across functionally equivalent networks in which identical outputs result for each input configuration

The attribution score of feature x_i for the network f is computationally expressed as the integral of gradients of f along the straight-line path from x to x':

$$\mathrm{IG}_i(x) \overset{\text{def}}{=} (x_i - x_i') \int_{\alpha=0}^{1} \frac{\partial f\left(x' + \alpha(x - x')\right)}{\partial x_i}$$

This formulation can be interpretated as the cumulative sensitivity of f to changes in the i^{th} feature. The integral is approximated numerically using Riemann summation.

Grad-CAM

Gradient-weighted Class Activation Mapping (*Grad-CAM*) [4] produces a coarse localization map, harnessing gradients of the network's output with respect to the feature maps, which are typically associated with the neurons in the last convolutional layer. This technique, an expansion of *Class Activation Mapping* (*CAM*), extends its applicability to a wide range of CNNs. The importance score for the $m \times n$ feature map A^k of a convolutional layer is calculated by averaging the gradient of the score y^c for class c with respect to A^k:

$$\alpha_k^c = \frac{1}{m \cdot n} \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{\partial y^c}{\partial A_{i,j}^k}$$

The resulting importance scores of every feature map are subsequently linearly combined and passed through a ReLU, as *Grad-CAM* only considers the positive influence of importance:

$$L_{\mathrm{Grad-CAM}}^{c} = ReLU\left(\sum_k \alpha_k^c A^k\right)$$

Finally, the activation maps are upsampled via bilinear interpolation to match the original input image resolution.

Method Comparison

Integrated Gradients and *DeepLIFT*, as shown in Table 1, have similar characteristics and produce comparable results, especially when using the rescale rule for *DeepLIFT*. However, *Integrated Gradients* is less efficient, since it needs to calculate the network's gradient for each Riemann sum component.

Grad-CAM avoids the need for a reference image and shows important feature maps instead of single pixels, which is often easier to interpret. In addition, *Grad-CAM* is class-discriminative, which allows visualization of only those features important to specific class decision (see Figure 1). Note that recently many extensions or combinations of the above methods have been developed, often yielding significant improvement.

Table 1: Comparison of the attribution methods DeepLIFT, Integrated Gradients *and* Grad-CAM.

Algorithm	DeepLIFT	Integrated Gradients	Grad-CAM
Application	Any artificial neural network	Any differentiable ML model	CNNs
Reference image required	Yes	Yes	No
Class-discriminative	No	No	Yes
Axioms	Sensitivity	Sensitivity, implementation invariance	None
Operating area	Global	Global	Local
Indicated attribution	Pixel-wise attributions; positive and negative attribution	Pixel-wise attributions; positive and negative attribution	Attribution of feature maps; only positive attribution

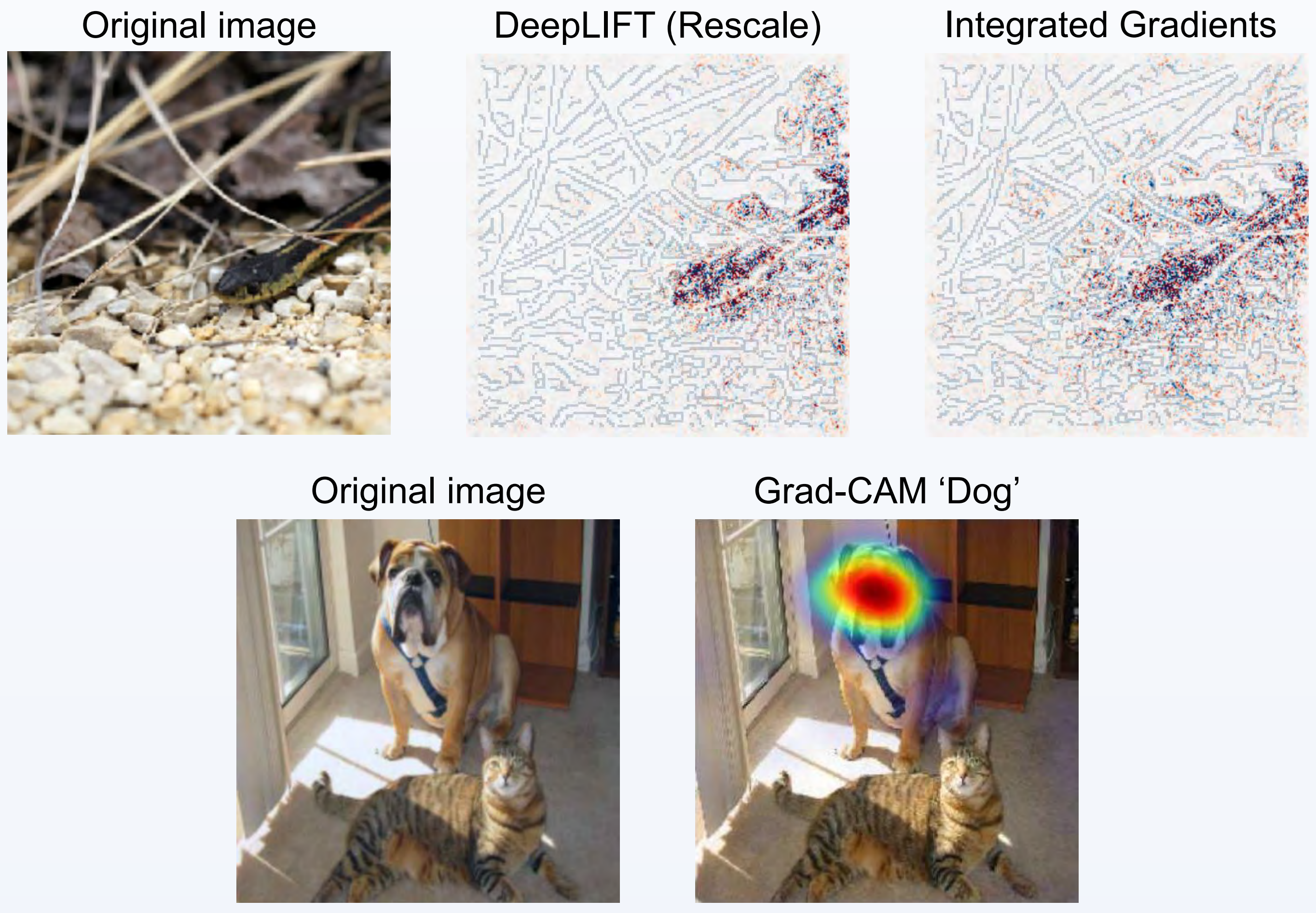

Figure 1: Feature importance detected by DeepLIFT, Integrated Gradients *(top) and* Grad-CAM *(bottom) for state-of-the-art CNNs [4],[5].*

Conclusions

- Highlighting relevant features is an important step in understanding the decision-making of typical black-box CNNs
- *DeepLIFT* determines the contributions of individual neurons, *Integrated Gradients* detects feature significance across the full range of input variables and *Grad-CAM* assigns class-specific activations to specific regions within images
- Integrated Gradients and *DeepLIFT* produce very similar results, but due to higher efficiency, *DeepLIFT* is usually preferable
- *Grad-CAM* generates a coarse localization map indicating important regions for a specific class decision

References

[1] G. Ras, N. Xie, M. Van Gerven, and D. Doran, 'Explainable deep learning: A field guide for the uninitiated', Journal of Artificial Intelligence Research, vol. 73, pp. 329–396, 2022.

[2] A. Shrikumar, P. Greenside, and A. Kundaje, 'Learning important features through propagating activation differences', in International conference on machine learning, 2017, pp. 3145–3153.

[3] M. Sundararajan, A. Taly, and Q. Yan, 'Axiomatic attribution for deep networks', in International conference on machine learning, 2017, pp. 3319–3328.

[4] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, 'Grad-cam: Visual explanations from deep networks via gradient-based localization', in Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.

[5] M. Ancona, E. Ceolini, C. Öztireli, and M. Gross, 'Towards better understanding of gradient-based attribution methods for deep neural networks', arXiv preprint arXiv:1711. 06104, 2017.