Visualizing and Understanding Deep Convolutional Neural Network in Image-based Structural Health Monitoring

Yuqing Gao and Khalid M. Mosalam

Department of Civil and Environmental Engineering, University of California Berkeley

Abstract

In this work, we emphasize the importance of understanding the trained deep convolutional neural network in the vision-based structural health monitoring (SHM) and investigate the applications of the recent widely-used visual interpretation methods, namely Guided backpropagation (GBP), Gradient-based class activation map (Grad-CAM) and Guided Grad-CAM, in the basic vision task of SHM, i.e., spalling detection.

Introduction

In the past decade, deep convolutional neural network (CNN) is increasingly attracting wide attention from researchers in many fields. It achieves the state-of-the-art performance and is proved to be far better than traditional methods. However, deep CNN is treated as a "black box," whose internal working principle is hard to understand by a human. This drawback significantly limits its usage in the real-world applications, e.g., vision-based structural health monitoring (SHM).

As an early work in vision-based SHM, we introduce Guided backpropagation (GBP), Gradient-based class activation map (Grad-CAM) and Guided Grad-CAM, to understand the network behavior through visualizing the backward signal or activated features. We conduct a comprehensive exploration towards the performance of these interpretation approaches in basic SHM vision tasks.

Signal Flow in Deep CNN

In the forward pass of CNN, F^l denotes the output from the l-th layer before activation function. A^l represents the activated feature maps from F^l, e.g., $A^l = \mathrm{ReLU}(F^l)$, where only positive entries in F^l are passed to A^l. S is the class score vector computed by the deep CNN, and S^c is one entry of it for a particular class c, $c \in \{1, 2, \dots, C\}$. In the backward pass, G^l denotes gradient of the output with respect to A^l in the l-th layer, i.e., $G^l = \frac{\partial S}{\partial A^l}$. When the signal further passes back through the activation function, the restoring gradient to bottom layers is denoted as R^l, where only positive gradient can be passed back. W_n^c represents the weight of the n-th feature map of A^l with respect to class c, as a scalar; $A^l_{i,j;n}$ is the entry with spatial location (i, j) in the n-th feature map of A^l. $\sum_{i,j} A^l_{i,j;n}$ is a shorthand notation of $\sum_i \sum_j A^l_{i,j;n}$ representing the global average pooling over a single feature map ignoring some constant multipliers.

$$R^l_{i,j} = \mathbb{1}(G^l_{i,j} > 0) \cdot G^l_{i,j} \tag{1}$$

$$\mathbb{1}(x) = \begin{cases} 0, & x < 0 \\ 1, & x \geq 0 \end{cases} \tag{2}$$

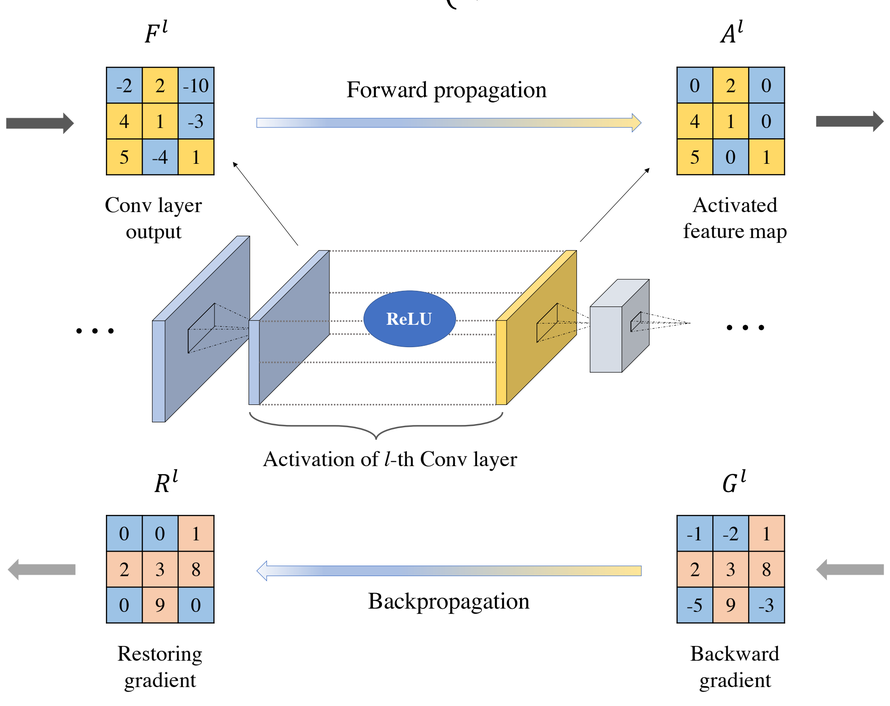

Figure 1 Forward and backward passes of one channel in the l-th Conv layer

Visual Interpretation Methodologies

Guided Backpropagation (GBP)

$$R^l_{i,j} = \mathbb{1}(F^l_{i,j} > 0) \cdot \mathbb{1}(G^l_{i,j} > 0) \cdot G^l_{i,j} \tag{3}$$

Gradient-based Class Activation Map (Grad-CAM)

$$\widetilde{W_n^c} = \frac{1}{M} \sum_i \sum_j \frac{\partial S^c}{\partial A^l_{i,j;n}} \tag{4}$$

$$L^c_{Grad-CAM} = \sum_n \widetilde{W_n^c} \cdot A^l_n \tag{5}$$

Guided Grad-CAM

Up-sampling $L^c_{Grad-CAM}$ to the image size, and further elementwise multiplying with the GBP saliency map.

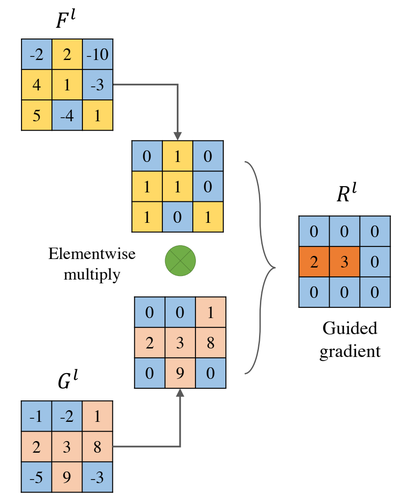

Figure 2 Formulation of restoring gradient in GBP

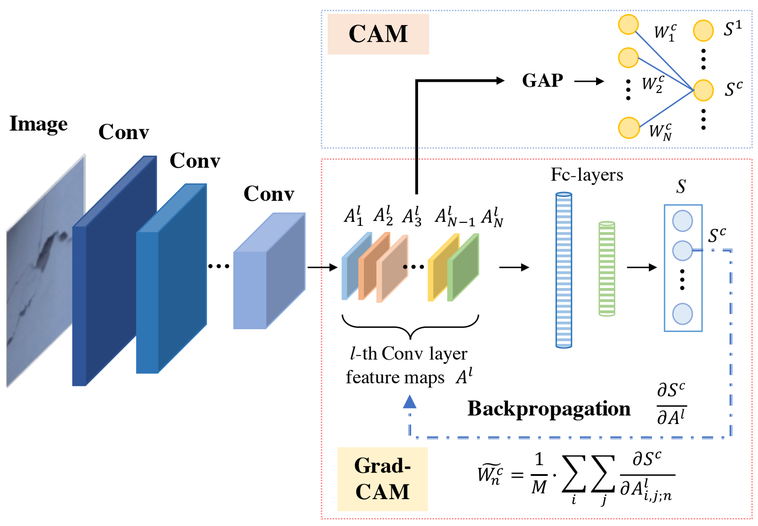

Figure 3 Weight computation in CAM and Grad-CAM

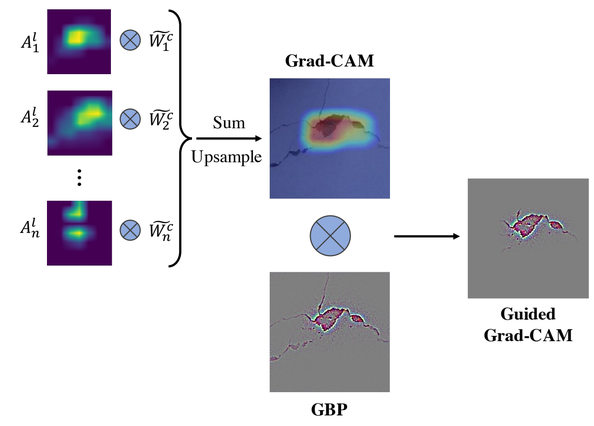

Figure 4 Saliency map of Guided Grad-CAM

Results & Visualization

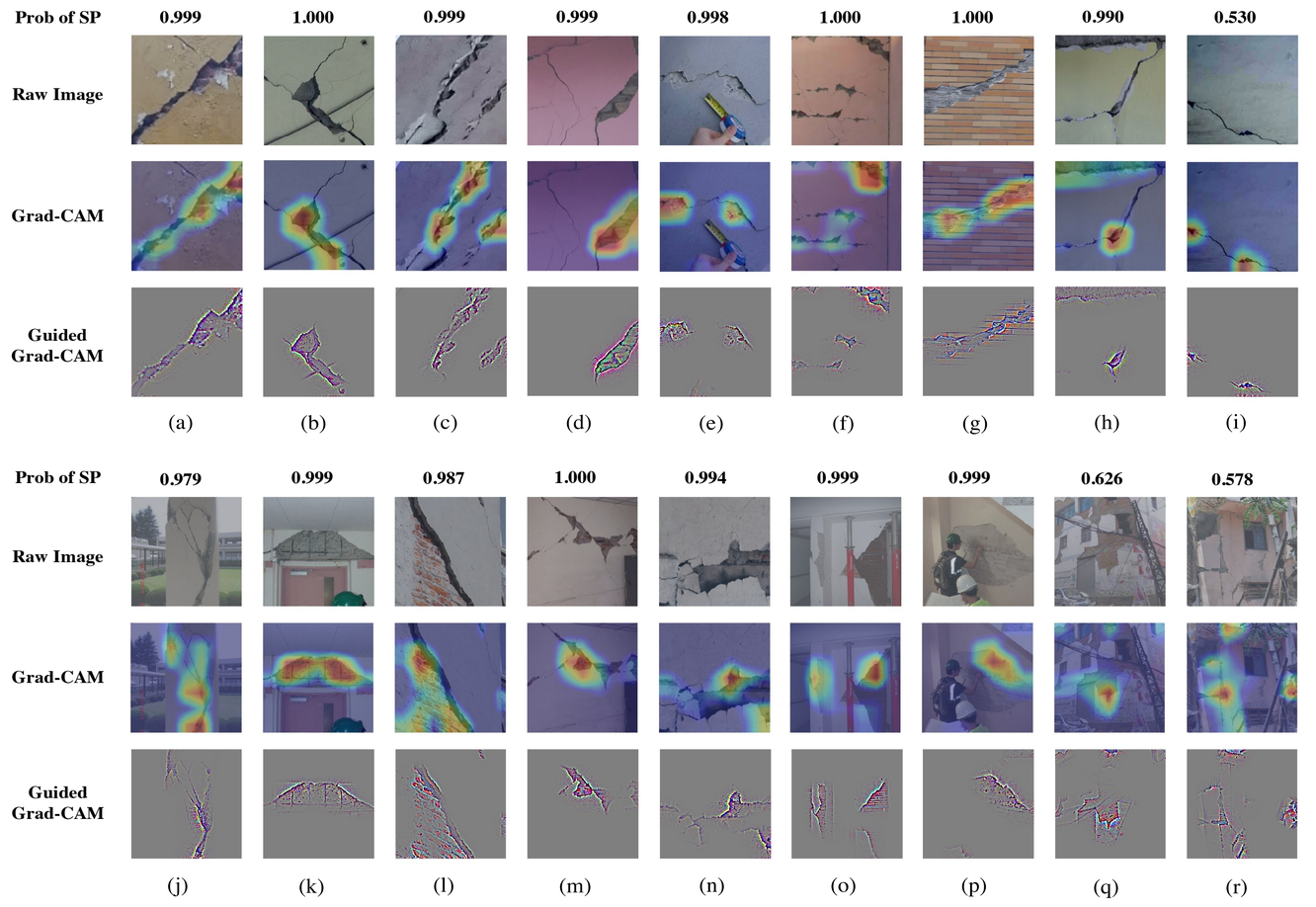

Figure 5 Saliency maps of Grad-CAM and Guided Grad-CAM in recognizing spalling

Acknowledgement

The author would like to acknowledge the funding support of TBSI.