Proceedings of the 33rd European Safety and Reliability Conference (ESREL 2023)
Edited by Mário P. Brito, Terje Aven, Piero Baraldi, Marko Čepin and Enrico Zio
 Published by Research Publishing, Singapore.
doi: 10.3850/978-981-18-8071-1_P290-cd

Segmenting without Annotating: Crack Segmentation and Monitoring via Post-hoc Classifier Explanations

Florent Forest, Hugo Porta, Devis Tuia, Olga Fink
EPFL, Switzerland. E-mail: {first}.{last}@epfl.ch

Monitoring the cracks in walls, roads and other types of infrastructure is essential to ensure the safety of a structure, and plays an important role in structural health monitoring. Automatic visual inspection allows an efficient, cost-effective and safe health monitoring, especially in hard-to-reach locations. To this aim, data-driven approaches based on machine learning have demonstrated their effectiveness, at the expense of annotating large sets of images for supervised training. Once a damage has been detected, one also needs to monitor the evolution of its severity, in order to trigger a timely maintenance operation and avoid any catastrophic consequence. This evaluation requires a precise segmentation of the damage. However, pixel-level annotation of images for segmentation is labor-intensive. On the other hand, labeling images for a classification task is relatively cheap in comparison. To circumvent the cost of annotating images for segmentation, recent works inspired by explainable AI (XAI) have proposed to use the post-hoc explanations of a classifier to obtain a segmentation of the input image. In this work, we study the application of XAI techniques to the detection and monitoring of cracks in masonry wall surfaces. We benchmark different post-hoc explainability methods in terms of segmentation quality and accuracy of the damage severity quantification (for example, the width of a crack), thus enabling timely decision-making.

Keywords: Crack detection, Image classification, Segmentation, Explainable AI, Attribution maps.

1. Introduction

The automated detection and segmentation of cracks in images is challenging due to the variety of crack aspects, the complexity and diversity of materials, and irregular illumination. Various approaches have been developed for this task, mainly based on image processing Yamaguchi et al. (2008); Hoang (2018). Data-driven approaches based on supervised learning have shown great performance, often using the popular U-Net neural network architecture Augustauskas and Lipnickas (2020). However, they require to label large amounts of images at pixel-level.

Post-hoc explainability methods aim at explaining the decisions of black-box models such as deep neural networks (see Arrieta et al. (2019) for a review). In particular, in this work, we focus on attribution methods, that associate a relevance to each feature in the input. The authors of Seibold et al. (2022) proposed to leverage the explanations of a classifier to segment damages in magnetic tiles and sewer pipe images, motivated by the fact that while annotating images for supervised segmentation is tedious, classification labels can be obtained at a fraction of the cost. In this work, we benchmark the ability of several post-hoc explainability methods, as well as post-processing steps, to generate high-quality segmentation masks for cracks in masonry building wall surfaces.

A major concern is the development and propagation of cracks over time, leading to increased stress and subsequent failure of the structure. The severity of a crack can be quantified, for instance, through width measurement Carrasco et al. (2021). Thus, we also study if these methods are usable to quantify damage severity and monitor its evolution, thus enabling timely decision-making.

2. From Classification to Segmentation

We propose the following methodology to generate crack segmentation masks:

(i) Train a binary classifier on positive (cracked) and negative (non-cracked) training images.
(ii) Perform inference on unseen test images. For each positive prediction, extract post-hoc explanations of the classifier and produce attribution maps for the positive class.
(iii) Post-process the resulting attribution maps:

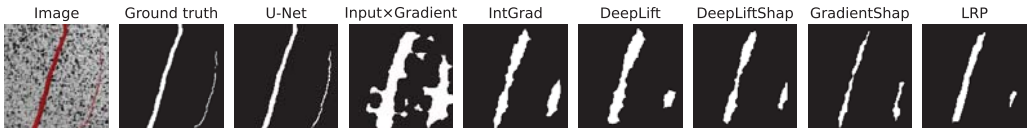

Fig. 1.: Visualization of crack segmentations obtained via post-hoc explainability methods and post-processing.

Table 1.: Crack segmentation quality for different explainability and post-processing methods (F1 score in %).

Post-processing thresh.	morph.	Baseline	Input×Grad	IntGrad	DeepLift	DeepLiftShap	GradientShap	LRP	U-Net (oracle)
simple	✗	12.18	14.64	18.91	22.58	24.03	13.88	22.17	83.67
	✓	4.73	23.30	27.74	34.44	**38.19**	20.61	**37.43**	
GMM	✗	18.05	21.54	28.92	31.70	37.07	19.78	28.39	
	✓	7.88	20.76	25.96	29.55	33.10	21.27	36.16	

(a) Thresholding using the *simple* or *GMM* strategies as in Seibold et al. (2022).

(b) Morphological closing and area opening operations, in order to close gaps in the mask and remove noisy attributions (see Figure 1).

We conducted experiments on the Experimental DIC (digital image correlation) cracks data set Rezaie et al. (2020), consisting in 256×256 image patches from stone masonry walls damaged in a shear-compression loading experiment. We complemented this data set with 874 additional negative patches coming from the same walls.

The crack classifier network is a VGG11 with 128 neurons in the fully-connected layers. In this study, we evaluated following post-hoc XAI techniques: Input×Gradient, Integrated Gradients (IntGrad), DeepLift, DeepLiftShap, GradientShap and Layer-wise Relevance Propagation (LRP). We also include a simple baseline where the image is just converted to gray-scale before the post-processing. As an oracle, we trained a U-Net11 on the segmentation labels of the training set.

The segmentation quality is evaluated by the F1 score on the test set. For each method, we report the results using combinations of thresholding and morphological post-processing in Table 1.

3. Crack Severity Monitoring

To assess the severity of cracks, we computed the number of cracks per patch (CPP) Pantoja-Rosero et al. (2022), the total crack area per patch, and the maximum crack width, using the width estimation method from Carrasco et al. (2021). We report the mean absolute error (MAE) or mean absolute percentage error (MAPE, in %) with the ground-truth measure for two of the methods in Table 2.

Table 2.: Crack severity assessment results.

Method	Post-processing thresh.	morph.	CPP MAE	Area MAPE	Width MAPE
DeepLift	simple	✓	0.81	146.0	264.6
	GMM	✓	**0.72**	352.1	374.2
LRP	simple	✓	0.90	**91.0**	**163.1**
	GMM	✓	0.82	261.8	257.6
U-Net (oracle)			0.74	20.1	20.8

References

Arrieta et al. (2019). Explainable Artificial Intelligence (XAI). arXiv:1910.10045.

Augustauskas, R. and A. Lipnickas (2020). Improved Pixel-Level Pavement-Defect Segmentation Using a Deep Autoencoder. *Sensors 20*(9).

Carrasco et al. (2021). Image-Based Automated Width Measurement of Surface Cracking. *Sensors 21*(22).

Hoang, N.-D. (2018). Detection of Surface Crack in Building Structures Using Image Processing Technique with an Improved Otsu Method for Image Thresholding. *Advances in Civil Engineering 2018*.

Pantoja-Rosero et al. (2022). TOPO-Loss for continuity-preserving crack detection using deep learning. *Construction and Building Materials 344*.

Rezaie et al. (2020). Comparison of crack segmentation using digital image correlation measurements and deep learning. *Construction and Building Materials*.

Seibold et al. (2022). From Explanations to Segmentation: Using Explainable AI for Image Segmentation. arXiv:2202.00315.

Yamaguchi et al. (2008). Image-Based Crack Detection for Real Concrete Surfaces. *IEEJ Transactions on Electrical and Electronic Engineering 3*(1).