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_P662-cd

Exploiting Explanations to Detect Misclassifications of Deep Learning Models in Power Grid Visual Inspection

Giovanni Floreale, Piero Baraldi

Energy Department, Politecnico di Milano, Milano, Italy. giovanni.floreale@polimi.it; piero.baraldi@polimi.it;

Enrico Zio

Energy Department, Politecnico di Milano, Milano, Italy & Mines Paris-PSL University, CRC, Sophia Antipolis, France. enrico.zio@polimi.it

Olga Fink

Intelligent Maintenance and Operations Systems, EPFL, Lausanne, Switzerland olga.fink@epfl.ch

In the context of automatic visual inspection of infrastructures by drones, Deep Learning (DL) models are used to automatically process images for fault diagnostics. While explainable Artificial Intelligence (AI) algorithms can provide explanations to assess whether the DL models focus on relevant and meaningful parts of the input, the task of examining all the explanations by domain experts can become exceedingly tedious, especially when dealing with a large number of captured images. In this work, we propose a novel framework to identify misclassifications of DL models by automatically processing the related explanations. The proposed framework comprises a supervised DL classifier, an explainable AI method and an anomaly detection algorithm that can distinguish between explanations generated by correctly classified images and those generated by misclassifications.

Keywords: black-box, explainable AI, explanations post-processing, fault diagnostics

1. Introduction

Inspection of infrastructures by drones is emerging as a viable option to monitoring their condition. By processing images collected by drones with Deep Learning (DL) models, the accuracy of detecting defects and degrading conditions in infrastructure components can be significantly improved. This, in turn, leads to increased efficiency in implementing targeted maintenance interventions. DL models are indeed capable of delivering good performance; however, they can be biased and may achieve accurate results by relying on non-causal shortcuts (Geirhos et al. 2020).

Different types of explainability techniques have been proposed to identify the elements of the input that contribute the most to the output of the DL model(Samek et al. 2021). With the help of these explanations, domain experts can identify instances where the DL model is making inferences based on irrelevant parts of the input, thereby indicating that the output should not be relied upon. However, the manual procedure of evaluating the explanations is inefficient and time-consuming, requiring each explanation to be analyzed by domain experts. In (Lapuschkin et al. 2019) the unsupervised spectral relevance analysis (SpRAy) method is proposed to semi-automatically post-process explanations. It has been shown to effectively identify clusters of similar explanations, thereby reducing the expert effort by focusing on the cluster prototypes alone. In this work, we propose a novel framework to automatically process explanations of a supervised DL model in order to identify its misclassifications and shortcuts.

2. Method

In this research, we automatically process the explanations to identify the unusual ones by applying a semi-supervised anomaly detection algorithm. In the proposed framework, after applying a supervised DL model, the images in the validation dataset are labelled as correctly or incorrectly classified by the DL model. Due to the high accuracy of DL models, the number of images labelled as incorrect is typically small, resulting in an imbalanced problem. In the subsequent step, explanations of images of the

validation dataset are obtained by applying CartoonX (Kolek et al. 2022a), an algorithm that generates relevance maps based on Rate Distortion Explanations (Kolek et al. 2022b). CartoonX extracts the most important features of the input in the wavelet domain and highlights them in the original image.

To process explanations, we apply the deep Semi-Supervised Anomaly Detection (*Deep* SAD) (Ruff et al. 2019) algorithm to generate a compact representation of explanations corresponding to correctly classified images, while also discerning explanations of incorrectly classified images. It is worth noting that in the proposed framework, a Deep SAD model is developed for each class of the classification problem. Finally, domain experts are requested to manually reclassify only those images whose explanations, as per the Deep SAD method, are dissimilar to typical explanations of correctly classified images.

3. Results

We evaluate the performance of the developed framework using a MobileNetV3 (Howard et al. 2019) DL model, whose task is to diagnose faults in insulators' shells based on images captured by drones. Table 1 presents the results obtained by applying *Deep* SAD to the images in the test dataset assigned to the class of broken insulators shells by the DL model. The performance metric considered is the classification accuracy, which represents the number of correctly classified images divided by the total number of images assigned by the DL model to the broken shell class. By reclassifying images with explanations dissimilar to those observed when the broken insulator shells are correctly classified, the classification accuracy increases from 89% to 95%. Domain experts are specifically asked to review only 20% of the test images, among which 36% are found to be incorrectly classified by the DL model. Additionally, among the remaining 64% of the images identified by *Deep* SAD as having explanations not similar to those of broken insulator shells, some shortcuts are discovered.

Table 1. Results of the application of Deep SAD to explanations.

Initial accuracy of DL model	89.20%
Final accuracy of DL model after revision	94.87%
Accuracy improvement of DL model	5.67%
Images fraction that is revised by experts	19.52%

Figure 1 illustrates an example image along with the corresponding CartoonX explanation. The explanation highlights the pixels that the DL model focused on obscuring the parts of the image that were not considered for classification. In this example, a shortcut is evident: the DL model correctly classifies the image as a broken insulator shell, but it mistakenly focuses on the background instead of the actual damage on the insulator shell. The consistent results obtained demonstrate that the proposed explanations-based framework effectively assists domain experts in identifying misclassifications and shortcuts of the DL model.

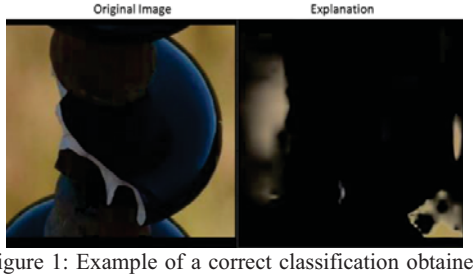

Figure 1: Example of a correct classification obtained by using a shortcut.

References

Geirhos, Robert et al. 2020. «Shortcut learning in deep neural networks». Nature Machine Intelligence 2 (11). doi:10.1038/s42256-020-00257-z.

Howard, Andrew et al. 2019. «Searching for mobileNetV3» in . Proceedings of the IEEE International Conference on Computer Vision. Libk. 2019-October. doi:10.1109/ICCV.2019.00140.

Kolek, Stefan et al. 2022a. «Cartoon Explanations of Image Classifiers» in . Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Libk. 13672 LNCS. doi:10.1007/978-3-031-19775-8_26.

———. 2022b. «A Rate-Distortion Framework for Explaining Black-Box Model Decisions» in . Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Libk. 13200 LNAI. doi:10.1007/978-3-031-04083-2_6.

Lapuschkin, Sebastian et al. 2019. «Unmasking Clever Hans predictors and assessing what machines really learn». Nature Communications 10 (1). doi:10.1038/s41467-019-08987-4.

Ruff, Lukas et al. 2019. «Deep Semi-Supervised Anomaly Detection» (ekainak 6). http://arxiv.org/abs/1906.02694.

Samek, Wojciech et al. 2021. «Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications». Proceedings of the IEEE 109 (3). doi:10.1109/JPROC.2021.3060483.