


Automation of Explainability Auditing for Image Recognition

Duleep Rathgamage Don, Kennesaw State University, USA*

Jonathan Boardman, Kennesaw State University, USA

Sudhashree Sayenju, Kennesaw State University, USA

Ramazan Aygun, Kennesaw State University, USA

 <https://orcid.org/0000-0001-7244-7475>

Yifan Zhang, Kennesaw State University, USA

Bill Franks, Kennesaw State University, USA

Sereres Johnston, The Travelers Companies, Inc., USA

George Lee, The Travelers Companies, Inc., USA

Dan Sullivan, The Travelers Companies, Inc., USA

Girish Modgil, The Travelers Companies, Inc., USA

ABSTRACT

XAI requires artificial intelligence systems to provide explanations for their decisions and actions for review. Nevertheless, for big data systems where decisions are made frequently, it is technically impossible to have an expert monitor every decision. To solve this problem, the authors propose an explainability auditing method for image recognition whether the explanations are relevant for the decision made by a black box model, and involve an expert as needed when explanations are doubtful. The explainability auditing system classifies explanations as weak or satisfactory using a local explainability model by analyzing the image segments that impacted the decision. This version of the proposed method uses LIME to generate the local explanations as superpixels. Then a bag of image patches is extracted from the superpixels to determine their texture and evaluate the local explanations. Using a rooftop image dataset, the authors show that 95.7% of the cases to be audited can be detected by the proposed method.

KEYWORDS

Classification, Computer Vision, Deep Learning, Explainability, Explainable Artificial Intelligence, Image Recognition

INTRODUCTION

During the last decade, artificial intelligence has claimed many achievements matching or surpassing human-level performance in some application domains such as object recognition. The performance of deep learning algorithms has been boosted with the introduction of additional layers or residuals

DOI: 10.4018/IJMD.332882

*Corresponding Author

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

from earlier layers continued to improve the performance (He et al., 2016). However, as the complexity of models has increased, the model interpretability has decreased, and as a result such black box models have become problematic in high-stakes decision-making domains, where safe and reliable performance are critical due to the high cost associated with errors (Guidotti et al., 2018). This is exacerbated by the realization that the patterns learned by discriminative deep architectures are less robust than what previously thought and vulnerability to adversarial attacks is the rule rather than the exception. In some cases, changing a single pixel is enough to fool a trained model (Su et al., 2019). Attacks can even be carried out in the real world by, for example, attaching a piece of black tape to a stop sign (Eykholt et al., 2018).

There are many ways we may wish to employ Explainable Artificial Intelligence (XAI) methods, and the choice of method and nature of the explanation should be informed by the problem context. Many different approaches to interpretability have emerged to meet this demand, and they can be categorized along several dimensions such as global vs. local, model-specific vs. model-agnostic, and intrinsic vs. post-hoc (Molnar et al., 2020; Rai, 2020). For deep neural networks, intrinsic interpretability may not be attainable. It has been noted that model interpretability and model flexibility or accuracy tend to be inversely related (Freitas, 2014). As the complexity of classification models increases, high accuracies in predictions can be achieved, but interpretability suffers. For example, Slack et al. (2019) investigate and conclude that decision trees and logistic regression are locally interpretable models while neural networks are not.

In contrast to global explainability techniques, which seek to explain the entire model (either by designing the model to be intrinsically interpretable or through an interpretable surrogate model), local explainability techniques provide explanations for individual predictions. Ribeiro et al., (2016) introduce local Interpretable Model-Agnostic Explanations (LIME) as a simple local explainability technique that generates simulated data points using random perturbations in the neighborhood of an instance to be predicted by the black-box model and fits a weighted linear regression on the simulated data to create explanations for the prediction. One of the main advantages of LIME is being model agnostic and hence, it may diminish the need for interpretable models. Usually, local explainability techniques provide interpretations of how an individual sample is analyzed, and the analysis may convince an expert to determine whether the model focuses on the right components or segments of data to make the decision. For example, Ribeiro et al. (2016), show that husky vs wolf image classification was done based on the signal of the background rather than focusing on the features of the animal. In other words, the learned model recognizes a domestic environment (e.g., home) compared to a wild environment (e.g., forest). This helps the expert determine whether the learned model is reliable or not, and this makes it a spectacular tool for individual analysis of samples. However, if the goal is to uncover systematic issues with the model, an expert must check the explanation of every sample.

Deep learning models may be trained on huge datasets of which the size may range from terabytes to petabytes. Monitoring explanations of these models by hand during the training process is out of the question. Even whenever it is possible, what matters is how those trained machine learning models behave in the wild for previously unseen data since critical decisions may rely on these models. Regardless of the possibility of manual checking, such a costly approach voids one of the main benefits of using machine learning—scalability.

This paper presents an automated explainability audit framework known as *ExplainabilityAudit* (DR Don et al., 2022) to investigate local interpretability in image recognition. As shown in Figure 1, the proposed method analyzes the reliability of classification by processing the explanations. After analyzing the explanations, it returns *satisfactory* if the explanations are good or *weak* if the explanations are poor. This technique requires training another model based on explanations. If this audit model determines that an explanation of a decision by the main model is weak (not reliable), this would require the involvement of a human expert to analyze the prediction and explanation. A human expert would only be required to step in for relatively few cases instead of potentially thousands or millions.

15 more pages are available in the full version of this document, which may be purchased using the "Add to Cart" button on the publisher's webpage: www.igi-global.com/article/automation-of-explainability-auditing-for-image-recognition/332882

Related Content

A Social Media Recommender System

Giancarlo Sperli, Flora Amato, Fabio Mercurio, Mario Mezzanzanica, Vincenzo Moscato and Antonio Picariello (2018). *International Journal of Multimedia Data Engineering and Management* (pp. 36-50).

www.irma-international.org/article/a-social-media-recommender-system/196248

The Future of WiMAX

Dennis Viehland and Sheenu Chawla (2011). *Handbook of Research on Mobility and Computing: Evolving Technologies and Ubiquitous Impacts* (pp. 314-326).

www.irma-international.org/chapter/future-wimax/50595

Learning by Playing: Development of an Interactive Biology Lab Simulation Platform for Educational Purposes

Vasilis Zafeiropoulos, Dimitris Kalles and Argyro Sgourou (2016). *Experimental Multimedia Systems for Interactivity and Strategic Innovation* (pp. 204-221).

www.irma-international.org/chapter/learning-by-playing/135131

A Combination of Spatial Pyramid and Inverted Index for Large-Scale Image Retrieval

Vinh-Tiep Nguyen, Thanh Duc Ngo, Minh-Triet Tran, Duy-Dinh Le and Duc Anh Duong (2015). *International Journal of Multimedia Data Engineering and Management* (pp. 37-51).

www.irma-international.org/article/a-combination-of-spatial-pyramid-and-inverted-index-for-large-scale-image-retrieval/130338

Emotion and Online Learning

Ileana Torres, Aubrey Statt and Kelly M. Torres (2022). *Online Distance Learning Course Design and Multimedia in E-Learning* (pp. 81-113).

www.irma-international.org/chapter/emotion-and-online-learning/299833