Deep Neural Network Interpretability Methods for Supervised and Unsupervised Problems

Andrzej Brodzicki[1][000000017713526X], Dariusz Kucharski[1][0000000201072407], Michał Piekarski[1,2][0000000193914263], Aleksander Kostuch[1][0000−0003−1242−9851], and Joanna Jaworek-Korjakowska[1][0000000301468652]

[1] Department of Automatic Control and Robotics, AGH UST, Krakow, Poland
{brodzicki,kucharski,piekarski,kostuch,jaworek}@agh.edu.pl
https://home.agh.edu.pl/ mdig/dokuwiki/doku.php
[2] SOLARIS National Synchrotron Radiation Centre, UJ, Krakow, Poland

Abstract. In recent years, deep neural networks (DNNs) have experienced a dynamic rise in applicability in many fields, from industry, through social media to healthcare. In this paper we focus on model interpretability for image analysis as it is a crucial point while deploying the methods in real life. We compare three visualisation algorithms including GradCAM, LIME and Occlusion that increase the model interpretability and check if the assessment is based on correct parts of the image or surrounding. We have compared the effectiveness of these methods in four different image processing research areas including 1) dermoscopic image classification, 2) lung nodule segmentation on CT scans, 3) classification of beam images for anomaly detection in synchrotron, 4) classification of seat occupancy. We briefly describe the model interpretability methods, compare achieved results and draw conclusions.

Keywords: Deep neural networks · Interpretability · Explainability · Supervised · Unsupervised

1 Introduction

One of the main advantage of deep learning is the ability to extract important features from raw data, with almost no preprocessing or expert knowledge. Such an approach is able to solve many computer vision tasks which couldn't be achieved with regular machine learning algorithms, however,this also leads to a low level o trustworthiness, as the user often does not understand how the network works and what criteria have been taken into account to draw the final decision. It is especially crucial in high-risk areas such as healthcare or autonomous driving. An answer to this is the idea of explainable and interpretable AI. The first term focuses on making users understand how it works. The second tries to trace the exact path of decision making inside the algorithm itself. Both terms are closely tied, as tools used for interpretability are often used for explanation too. This paper has been organized as follows: in Section 2 we provide a brief description of three state-of-the-art interpretability techniques from the

field of computer vision, in Section 3 we present the achieved outcomes in different computer vision problems including synchrotron beam stability assessment or seat occupancy classification in vehicle interior.

2 Deep learning model interpretability methods

2.1 Related works

Q. Xhang et al. described four groups for interpretability of convolutional neural networks (CNNs): filters visualisation as a most direct way to explore patterns hidden inside a neural unit, pattern retrieval based on extraction of mid-level features (conv-layers), model diagnosis by checking image regions accountable for the prediction (i.e. gradient-based methods, LIME) and finally, learning a more meaningful representation right from its design [8]. Hedstrm et al. introduced a versatile tool for quantification of a wide range of evaluation metrics [1].

2.2 Interpretability algorithms description

We have chosen three popular model interpretability algorithms which are mostly based on visualization of CNNs activities: GradCAM, LIME and Occlusion. Gradient-weighted Class Activation Mapping (GradCAM) utilizes the gradients of a target (i.e image of a certain class in a classification network) fed into the last convolutional layer to produce a map highlighting the important regions in the image for predicting this particular class [6]. Local Interpretable Model-agnostic Explanations (LIME) explains the predictions of a classifier by learning an interpretable model locally around the prediction [5]. Finally, Occlusion map is performed by analysing the classifier output by occluding portions of the input image, showing which parts of the scene are important for classification [7].

3 Experimental comparison of interpretability methods

3.1 Melanoma malignancy classification in different anatomic sites

Diagnosing a melanoma (a deadly skin cancer) is a challenging task even for expert dermatologists. There is a need for decision-making systems to assess the variety of morphological arrangements in skin moles. In paper [2] we focused on classifying skin lesions originating in different anatomic sites of the body, with a 97% ACC. We also proposed a statistical metric, based on the overlap of Grad-CAM heatmaps and the segmentation ground-truth, to quantify the interpretability (Fig. 1).

3.2 Nodule detection on patches extracted from lungs CT scan

The objective of this research is to localize a nodule change on lungs CT scan images. First step of an algorithm is to extract patches from a single scan and

asses if they contain a nodule change. Fig 1 shows that trained neural network is focusing on finding a nodule change on such patches - GradCAM and Lime show network's attention focused on a nodule itself and in occlusion case one can notice negative predictions if areas around nodule were occluded.

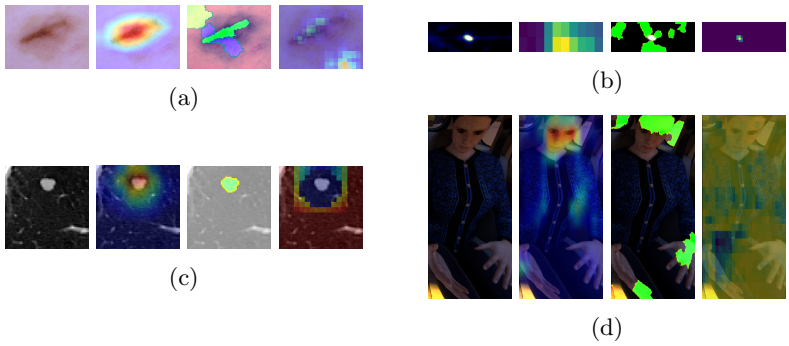

Fig. 1: DNN attention areas obtained from interpretability algorithms - from left to right: input image, GradCAM, LIME and Occlusion results for: a) melanoma classification, b) anomaly detection in synchrotron, c) lung nodule detection, d) seat occupancy classification. The more intense the color, the greater the attention of the neural network.

3.3 Electron beam anomaly detection in Pinhole diagnostic line

Detection of anomalies and instability in diagnostic signals in the SOLARIS synchrotron, with particular emphasis on the images of transverse electron beam profile from the Pinhole diagnostic beamline [4], allows operator to better tune crucial elements of the storage ring (i.e 3rd harmonic cavities) and to observe the state of the entire system. In such complicated and distributed systems, detecting unwanted events and understanding them prevents financial losses, unplanned downtime and damage to the infrastructure. Therefore it is crucial to build interpretable models. Our model results (94% ACC) are shown in Fig. 1.

3.4 Seat occupancy classification in vehicle interior

Classification of seat occupancy in in-vehicle interior is a promising area in new generation cars. In our study [3], we provided an interpretable solution (using ResNet, DenseNet, EfficientNet) that identify object parts without direct supervision. We also proposed two new statistical metrics based on the multivariate Gaussian distribution in order to assess heatmaps without using human-labeled objects. We demonstrated that our interpretability results correlate with the accuracy and can be implemented to work with any resolution for various applications (Fig. 1). Extensive experiments were carried out on 7,500 BMW X5 images from SVIRO database. The model achieved a state-of-the-art result (79.87%

ACC, 95.92% recall, 90.32% spec.) in the domain of seat occupancy classification into seven main categories: empty infant in infant seat, child in child seat, adult, everyday object, empty infant seat and empty child seat.

4 Conclusions

In this work, we have proved importance of the interpretability of the DNN models by testing and comparing different methods in various tasks. During the analysis, most of the results were in line with expectations, but we observed individual cases where the concentration of the network falls on the area that is actually the background for a significant object. With this knowledge, we can consider such a result not entirely reliable, even if it is positive, which allows us to improve the operation of the model by supplementing the data set with similar examples. In summary, the conclusions drawn from the tested solutions allow not only to understand and explain the operation of the models, but also to improve their final reliability. Further research efforts will in the first place be directed to the integration of various methods of interpretability in order to find a method that is even more versatile in terms of the problem being solved.

Acknowledgments

We gratefully acknowledge the funding support of the "Excellence initiative—research university" programme for the AGH UST.

References

1. Hedström, A., et al.: Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations. ArXiv (2022)
2. Jaworek-Korjakowska, J., Brodzicki, A., Cassidy, B., Kendrick, C., Yap, M.H.: Interpretability of a deep learning based approach for the classification of skin lesions into main anatomic body sites. Cancers **13** (2021)
3. Jaworek-Korjakowska, J., Kostuch, A., Skruch, P.: Safeso: Interpretable and explainable deep learning approach for seat occupancy classification in vehicle interior. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2021)
4. Kisiel, A., Marendziak, A., Ptaszkiewicz, M., Wawrzyniak, A.: X-ray pinhole camera for emittance measurement in solaris storage ring (05 2019)
5. Ribeiro, M.T., et al.: Why should i trust you?: Explaining the predictions of any classifier. ACM KDD: Knowledge Discovery and Data Mining (2016)
6. Selvaraju, R.R., et al.: Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization (2016)
7. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: ECCV (2014)
8. Zhang, Q., Wu, Y.N., Zhu, S.C.: Interpretable convolutional neural networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018)