

Evaluation of Explainable AI methods for Classification Tasks in Visual Inspection

Björn Forcher^{1,*†}, Patrick Menold^{1†}, Moritz Weixler[†], Jörg Schmitt^{1†} and Samuel Wagner^{1†}

¹Robert Bosch GmbH, Wernerstraße 51, 70469 Stuttgart, Germany

Abstract

Methods of the eXplainable Artificial Intelligence (XAI) gain more and more interest in the machine learning (ML) community. For explaining neural networks, a lot of methods have been proposed, especially in the context of computer vision (CV). These approaches aim at explaining the decisions by means of sensitivity or importance of input features. In this paper, an application in the field of visual inspection (VI) in the manufacturing domain is analyzed. As different XAI methods produce interpretations of varying quality, we propose a metrics bundle to value the quality of those algorithms, e.g. Gradient or Guided Backpropagation. The bundle includes a new approach of measuring the correctness of the explanation and enables developers to rely on the most appropriate method for their use cases.

Keywords

Evaluation, Metrics, Explainable AI, Visual inspection

1. Introduction

As the scientific field of artificial intelligence is evolving rapidly in the past few years, concerns about the safety, security, reliability, and resiliency of these systems are growing. There have been increasing efforts to understand ML models in order to detect weaknesses (e.g. correlated features) at an early stage. These methods are known collectively under the term XAI [1] and are used to explain the decisions of an AI system.

In this paper, an application of visual inspection in the manufacturing of fuel injection equipment (FIE) systems is in focus. We implemented a neural network which is used to detect coating failures in FIE components. For improving the model we applied various XAI methods producing explanations of varying quality. In order to evaluate these methods objectively, a metric composition is proposed. It quantifies the ability of the XAI methods to accurately describe the sensitivity of the model at the given sample as correct as possible, the ability to compensate noise in the model function and the computational speed calculating the XAI algorithm.

Late-breaking work, Demos and Doctoral Consortium, colocated with The 1st World Conference on eXplainable Artificial Intelligence: July 26–28, 2023, Lisbon, Portugal


*Corresponding author.

†These authors contributed equally.

✉ bjoern.forcher@de.bosch.com (B. Forcher); patrick.menold@de.bosch.com (P. Menold); joerg.schmitt@de.bosch.com (J. Schmitt); samuel.wagner@de.bosch.com (S. Wagner)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

This paper is structured as follows. In the next section we present the applied XAI methods and the proposed metric composition. Section 3 shows the coating failure use case, the applied XAI methods and our metrics bundle. The final section summarizes our findings and gives an outlook of further investigations.

2. XAI methods and metrics

There are many XAI methods which are used to provide interpretations for ML models (see [2], [3], [4] or [5]). They can be classified by various characteristics such as agnosticism or locality [6]. Local approaches are mainly based on one example of the data set and reveal how features contribute to the output. Global approaches on the other side take the whole data set into account [7]. The agnosticism characteristic describes whether the approach works only for specific ML models (model-specific) or if it can be applied to any model (model-agnostic) [8]. In this work, we focus on model-specific methods for neural networks which provide local interpretations, namely Gradient Backpropagation (GrB) [9], Deconvolutional Networks (DeCon) [10] and Guided Backpropagation (GuB) [11].

The main question is how to decide which method provides good explanations [12]. Regarding our use case explanations should be correct and not susceptible to model noise. In addition, fast computation time is important due to real-time application in assembly lines (compare also to [13]). An explanation is correct if it directly represents, how the model made its decision (compare to [14] or [15]). This aspect can be determined by the sensitivity of a model which characterizes the behavior of infinitesimal changes in the input. Gradient methods, such as the above-mentioned ones, can be leveraged for that. In this context, the term feature sensitivity describes the ability of an XAI method to determine the sensitivity of a model as correct as possible. The second property is to be free of model-induced noise. The ability of a method to reduce the influence of model-induced noise is called noise susceptibility. In the following we provide a brief description of the metrics (see [16] for more details).

As mentioned above, **Feature Sensitivity** takes only the local context of the model function into account. Let $f_c(x_t)$ be the prediction probability for the testing input x_t and let x_p be a small deviation vector. We define the feature sensitivity metric M_s as follows.

$$M_s = \frac{1 - \frac{f_c(x_p)}{f_c(x_t)}}{\alpha}$$

The deviation vector x_p is based on the attributions of x_t . The attributions A are a general measure of the contribution of a single input value to the predication (compare to [17]). The deviation vector x_p is now chosen sample wise as percentage α using the L_2 norm and is defined as follows:

$$x_p = x_t - \frac{\alpha \cdot \|x_t\|_2}{\|A\|_2} \cdot A$$

There are many methods to calculate a sensitivity score (see [18]). In our approach we use the gradient-based attributions A to derive x_p and scale M_s to a range from -1 to 1 . For this purpose the score is divided by the score of the exact gradient calculated using GrB. This

automatically gives GrB has a score of 1 as the ideal algorithm. To get rid of the influence of arbitrarily chosen samples, the score can be calculated over as many samples as possible. For any method the scores of all samples form a stochastic distribution. As a single score the mean of all sample scores is used.

To find a metric to account for **Noise Susceptibility**, a definition of noise is needed. For any model trained on a finite amount of samples, the model function can only classify these samples exactly. The difference between an imaginary ideal model function trained on infinite samples and the present model function can be considered as an error. For this metric the error is assumed to be dominated by high frequency noise terms. For any XAI method, the attributions are as well assumed to consist of the ideal part A_{ideal} and the error A_ϵ introduced by the noise $A = A_{ideal} + A_\epsilon$. According to Balduzzi et al. [19], the gradient of the model function is even more susceptible to noise than the model function itself. The noise of the gradients is usually high frequency, especially for deep networks. The attribution error A_ϵ is therefore also expected to be large. This contrasts to the low frequent ideal attributions A_{ideal} . By applying a low pass image filter on the attribution map A the ideal attributions A_{ideal} are reconstructed. For the present application a Gaussian filter is used. The attribution error A_ϵ is assumed to be low for a method that is not susceptible to model noise and high for a method vulnerable to model induced noise. To quantitatively evaluate the model induced noise the structural similarity (SSIM) [20] is used to compare the attribution mask A before and after filtering. For the SSIM two images x and y are compared based on their difference in luminescence $l(x, y)$, contrast $c(x, y)$ and structure $s(x, y)$. The SSIM should yield a maximum score of 1 if (and only if) the attributions are identical. The SSIM is calculated window-wise for a number of local windows of the images. The final noise score M_N uses the mean SSIM over all windows.

The metric **Computational Speed** M_C is determined by the computed samples per second and is normalized by means of the fastest algorithm. Hence, M_C ranges between 0 and 1 whereas value 1 identifies the fastest algorithm.

3. Applying XAI Methods and Metrics

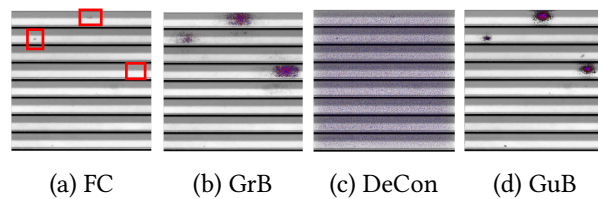
Visual inspection is an important application in the manufacturing domain. Pictures are taken in certain production steps in order to detect defective parts. In our use case we need to recognize coating failures on a cylindrical part. The ring showing the coating is extracted from the raw image, unrolled, straightened and stacked in lines showing 45° each. The model aims to identify not only good parts (OK), but also to distinguish between four different failure types (BLACK, DAMAGE, SCRATCH and SILVER).

The used ML model applies a modified ResNet50 architecture (see [21]). The original ResNet50 network is scaled up by a factor of 4 in width and height, respectively. Instead of $224 \times 224 \times 3$ the network uses input images of size $896 \times 896 \times 3$. All convolutional layers are also scaled up by the same factor. To be able to use the original fully connected layers at the end of the network without re-scaling, an average pooling layer is introduced.

The samples for bad classes are chosen equally. Here, the scores are calculated per class and for every considered XAI method an average is build. For the feature sensitivity the deviation percentage α is chosen empirically as 0.00001. The result can be seen in the following table.

Method	M_S	M_N	M_C
Gradient Backpropagation (GrB)	1.000	0.668	1.000
Deconvolutional Network (DeCon)	-0.050	0.358	0.586
Guided Backpropagation (GuB)	0.000	0.880	0.549

The generated sensitivity show that DeCon and GuB loose quite a lot of their correctness when identifying sensitive features. In addition, DeCon shows bad results regarding noise and speed and thus, it could be excluded for the coating failure use case. Regarding GrB and GuB a clear preference could not be derived. GrB is the fastest algorithm but GuB shows a better performance with respect to noise. The figure below illustrates our findings. The most left image represents a faulty case (FC) containing three problematic areas. The figure reveals that DeCon provides bad results for this use case. GrB and GuB on the contrary highlight these areas correctly.



4. Conclusion

GrB and GuB are useful for interpreting the coating failure use case. All anomalies get highlighted. DeCon does not work properly for this use case and should not be used here.

However, the sensitivity score can only be applied to gradient-based explanation methods. To include other XAI methods such as LRP [22], DeepLift [23], GradCam [24] or GradCam++ [25] which do not compute sensitivity scores but rather relevance scores, a clearer definition of correctness is needed with respect to all applicable XAI methods. Simply describing correctness as the ability to choose important attributions is not sufficient. Possibly, a combination of existing and new metrics could lead to clearer results. Also more metrics are needed to account for problems of different methods. For example, Guided Backpropagation is invariant to model randomization (see Adebayo et al. [26]).

Acknowledgments

Thanks to the developers of ACM consolidated LaTeX styles <https://github.com/borisveytsman/acmart> and to the developers of Elsevier updated L^AT_EX templates <https://www.ctan.org/tex-archive/macros/latex/contrib/els-cas-templates>.

References

- [1] M. van Lent, W. Fisher, M. Mancuso, An explainable artificial intelligence system for small-unit tactical behavior, in: Proceedings of the 16th Conference on Innovative Applications of Artificial Intelligence, IAAI'04, AAAI Press, 2004, p. 900–907.
- [2] A. Holzinger, A. Saranti, C. Molnar, P. Biecek, W. Samek, Explainable AI Methods - A Brief Overview, Springer International Publishing, 2022, pp. 13–38.
- [3] P. Linardatos, V. Papastefanopoulos, S. Kotsiantis, Explainable ai: A review of machine learning interpretability methods, *Entropy* 23 (2021).
- [4] B. K. Iwana, R. Kuroki, S. Uchida, Explaining convolutional neural networks using softmax gradient layer-wise relevance propagation, in: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), 2019, pp. 4176–4185.
- [5] E. Tjoa, C. Guan, A survey on explainable artificial intelligence (XAI): Toward medical XAI, *IEEE Transactions on Neural Networks and Learning Systems* 32 (2021) 4793–4813.
- [6] C. Molnar, *Interpretable Machine Learning - A Guide for Making Black Box AI Explainable*, Independently published, 2022.
- [7] U. Bhatt, A. Xiang, S. Sharma, A. Weller, A. Taly, Y. Jia, J. Ghosh, R. Puri, J. M. F. Moura, P. Eckersley, Explainable machine learning in deployment, in: In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020. [arXiv:1909.06342](https://arxiv.org/abs/1909.06342).
- [8] A. Carrillo, L. F. Cantú, A. Noriega, Individual explanations in machine learning models: A survey for practitioners, 2021. [arXiv:2104.04144](https://arxiv.org/abs/2104.04144).
- [9] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, in: Y. Bengio, Y. LeCun (Eds.), 2nd International Conference on Learning Representations (ICLR), 2014. [arXiv:1312.6034](https://arxiv.org/abs/1312.6034).
- [10] M. D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, 2013. [arXiv:1311.2901](https://arxiv.org/abs/1311.2901).
- [11] J. T. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller, Striving for simplicity: The all convolutional net, 2015. [arXiv:1412.6806](https://arxiv.org/abs/1412.6806).
- [12] R. R. Hoffman, S. T. Mueller, G. Klein, J. Litman, Metrics for explainable ai: Challenges and prospects, *ArXiv abs/1812.04608* (2018).
- [13] I. Kakogeorgiou, K. Karantzalos, Evaluating explainable artificial intelligence methods for multi-label deep learning classification tasks in remote sensing, *International Journal of Applied Earth Observation and Geoinformation* 103 (2021) 102520.
- [14] G. K. Santhanam, A. Alami-Idrissi, N. Mota, A. Schumann, I. Giurgiu, On evaluating explainability algorithms, 2019.
- [15] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. van Keulen, C. Seifert, From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai, *ACM Comput. Surv.* (2023). Just Accepted.
- [16] M. Weixler, Validation of Machine Learning Models with Algorithms from the Area of Explainable AI for Classification and Regression Tasks, Master's thesis, University Stuttgart, 2021.
- [17] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, 2017. [arXiv:1703.01365](https://arxiv.org/abs/1703.01365).
- [18] C. Yeh, C. Hsieh, A. S. Suggala, D. I. Inouye, P. Ravikumar, How sensitive are sensitivity-

- based explanations?, CoRR abs/1901.09392 (2019). URL: <http://arxiv.org/abs/1901.09392>. arXiv:1901.09392.
- [19] D. Balduzzi, M. Frean, L. Leary, J. Lewis, K. W.-D. Ma, B. McWilliams, The shattered gradients problem: If resnets are the answer, then what is the question?, 2018. arXiv:1702.08591.
- [20] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, Image quality assessment: from error visibility to structural similarity., *IEEE Transactions on Image Processing* 13 (2004) 600–612.
- [21] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, 2015. arXiv:1502.01852.
- [22] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, *PLOS ONE* 10 (2015) 1–46. URL: <https://doi.org/10.1371/journal.pone.0130140>. doi:10.1371/journal.pone.0130140.
- [23] A. Shrikumar, P. Greenside, A. Kundaje, Learning important features through propagating activation differences, 2019. arXiv:1704.02685.
- [24] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, *International Journal of Computer Vision* 128 (2019) 336–359.
- [25] A. Chattopadhyay, A. Sarkar, P. Howlader, V. N. Balasubramanian, Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks, in: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 2018, pp. 839–847.
- [26] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, B. Kim, Sanity checks for saliency maps, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, volume 31, Curran Associates, Inc., 2018.