Highlighting Embeddings' Features Relevance Attribution on Activation Maps

Jože M. Rožanec
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia
joze.rozanec@ijs.si

Erik Koehorst
Philips Consumer Lifestyle BV
Drachten, The Netherlands
Erik.Koehorst@philips.com

Dunja Mladenić
Jožef Stefan Institute
Ljubljana, Slovenia
dunja.mladenic@ijs.si

ABSTRACT

The increasing adoption of artificial intelligence requires a better understanding of the underlying factors affecting a particular forecast to enable responsible decision-making and provide a ground for enhancing the machine learning model. The advent of deep learning has enabled super-human classification performance and eliminated the need for tedious manual feature engineering. Furthermore, pre-trained models have democratized access to deep learning and are frequently used for feature extraction. Nevertheless, while much research is invested into creating explanations for deep learning models, less attention was devoted to how to explain the classification outcomes of a model leveraging embeddings from a pre-trained model. This research focuses on image classification and proposes a simple method to visualize which parts of the image were considered by the subset of the most relevant features for a particular forecast. Furthermore, multiple variants are provided to contrast relevant features from a machine learning classifier and selected features during a feature selection process. The research was performed on a real-world dataset provided by domain experts from *Philips Consumer Lifestyle BV*.

KEYWORDS

explainable artificial intelligence, feature importance, activation map, GradCAM, image classification, smart manufacturing, defect detection

1 INTRODUCTION

The increasing adoption of artificial intelligence has posed new challenges, including enforcing measures to protect the human person from risks inherent to artificial intelligence systems. One step in this direction is the European AI Act [12], which considers that different artificial intelligence systems must conform to a different set of requirements according to their risk level, linked to the particular domain and potential impact on health, safety, or fundamental rights [15]. In this context, explainable artificial intelligence, a sub-field of machine learning, has gained renewed attention with the advent of modern deep learning [22], given that it researches how more transparency can be brought to opaque machine learning models. While transparency in the regulatory context is sought to enable responsible decision-making, it provides valuable insights to enhance the workings of machine learning models, too.

Information Society 2023, 9–13 October 2023, Ljubljana, Slovenia

The field of explainable artificial intelligence can be traced back to the 1970s [18]. A key question posed by the researchers is what makes a good explanation. Arrieta et al. [2] consider that a good explanation must take into account at least three elements: (a) the reasons for a given model output (e.g., features and their value ranges), (b) the context (e.g., context on which inference is performed), and (c) how are (a) and (b) conveyed to the target audience (e.g., what information can be disclosed and the vocabulary used, among others). When considering images, maps frequently present explanations that contrast particular model information on top of the original input image (e.g., saliency maps, activation maps, heat maps, or anomaly maps [13, 24]). Other approaches can be extracting and highlighting super-pixels relevant to a specific class [16] or the occlusion of background parts irrelevant to the model. Such outputs convey (a) the reasons for a given model output by highlighting the images, (b) the context on which inference is performed (by overlaying the information on top of the image used for inference), and (c) using an agreed approach to convey to the user what is considered more relevant and what is not.

Multiple approaches have been developed to explain the inner workings of image classifiers. LIME (Local Interpretable Model-Agnostic Explanations) [16] approached this challenge by retrieving predicted labels for a particular class and showing the segmented superpixels that match each class. GradCAM[19] has taken another approach and created activation maps considering the weight of the activations at particular deep learning model layers by the average gradient. Many approaches were developed afterward, following the same rationale. For example, GradCAM++[3], XGradCAM[9], or HiResCAM[6] work like GradCAM but consider second-order gradients, scale the gradients by the normalized activations, or element-wise multiply the activations with the gradients respectively. Other possible approaches are leveraging insights resulting from image perturbation [8] or methods that acquire and display samples similar or counterfactual to the predicted instance [4, 17].

The development of information and communications technologies fostered the emergence of the Industry 4.0 paradigm as a technology framework to integrate and extend manufacturing processes [23]. In this context, the increasing adoption of artificial intelligence enables greater automation of manufacturing processes such as defect inspection [7] and urges the adoption of explainable artificial intelligence to develop users' trust in the models and foster responsible decision-making based on the insights obtained regarding the underlying machine learning model [1].

From the literature mentioned above and several surveys on this topic [5, 13, 14, 17, 20, 21], it was found that the authors did not contemplate how explanations can be provided in scenarios where feature embeddings are extracted with a deep learning model and then used to train a separate machine learning model.

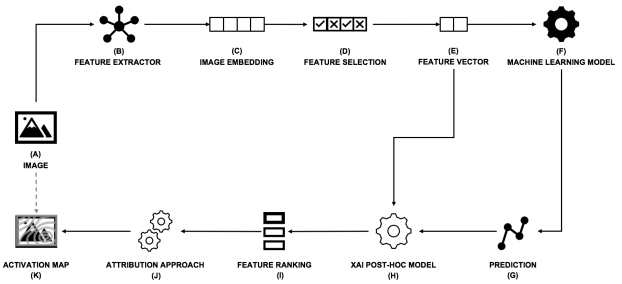

Figure 1: To classify an image, a feature extractor is used to create an embedding, from which certain values are extracted to create a feature vector. The machine learning model issues a prediction, which, along with the feature vector, is used to create a feature ranking. The attribution approach considers the highest-ranking features to generate an activation map.

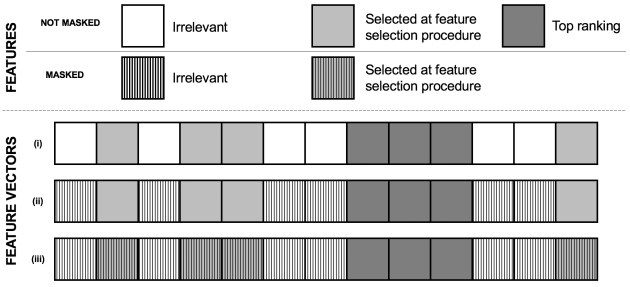

Figure 2: Given an image embedding (i), we can mask it to display (ii) features selected at the feature selection procedure (including the top ranking classifier's features, or (iii) can mask it to display only the top ranking classifier's features.

Figure 3: Sample images from the dataset provided by *Philips Consumer Lifestyle BV*. Three categories are distinguished: images corresponding to non-defective items (good) and images corresponding to two defect types (double-printed and with interrupted prints).

The present research addresses this void by proposing an unsupervised approach to generate activation maps based on the feature ranking obtained for a particular forecast. The research is performed on a real-world dataset provided by *Philips Consumer Lifestyle BV* and related to defect inspection.

This paper is organized as follows. First, section 2 describes the explainability approach developed and tested in this research. Section 3 describes the experiments performed to assess different value imputation strategies, and Section 4 informs and discusses the results obtained. Finally, Section 5 concludes and describes future work.

2 HIGHLIGHTING EMBEDDINGS' FEATURES RELEVANCE ATTRIBUTION ON ACTIVATION MAPS

The increasing amount of pre-trained deep learning models make them the default choice for feature extraction when working with machine learning models for images. Nevertheless, the disconnect between the machine learning model built on top and the deep learning model used to extract the image embedding makes it challenging to provide good explanations to the user. This research proposes an approach to bridge the gap (see Fig. 1). In particular, we leverage the fact that similar images or fragments of images result in embeddings or parts of embeddings that are close to each other. This property can be exploited when building activation maps, computing the similarity between a reference image (e.g., the image of a horse) and the image under consideration to find where such class can be found in the image under consideration (e.g., given the image of a farm, highlight where the horses are located). Nevertheless, if instead of using some reference image, the image that is an input to the machine learning model is leveraged as a reference, (i) no noise is introduced due to the dissimilarity of the images, and (ii) no beforehand knowledge regarding the classes of interest is required. Therefore, a key issue must be resolved: how do both embeddings differ to ensure that such difference is exploited to build an activation map?

Two options are envisioned in this research (see Fig. 2): given (i) the image embedding, two variations can be considered for value imputation: (ii) mask all the values in the embedding except for the ones corresponding to top-ranking features, (iii) mask all the values in the embedding except for the ones corresponding to selected features and top-ranking features, using different values for each of them. By doing so, the highest similarity in the image will be found in regions related to top-ranking features or selected features. Considering selected and top-ranking features provides additional insights into what information was provided to the model and what information was considered the most important by the model. These two approaches are explored in Section 3.

3 EXPERIMENTS

We experimented with a real-world dataset of logos printed on shavers provided by *Philips Consumer Lifestyle BV*. The dataset consisted of 3518 images considered within three categories (see Fig. 3): non-defective images and images with two kinds of defects (double-printed logos and interrupted prints). To extract features from the images, the ResNet-18 model [10] was used, extracting the features before the fully connected layer. Mutual information was used to evaluate the most relevant features and select the *top K*, with $K = \sqrt{N}$, where N is the number of data instances in the train set, as suggested in [11]. The dataset was divided into train (75%) and test (25%), and a random forest classifier was trained on it, achieving an AUC ROC (one-vs-rest) score of 0.9022.

Three images from the test set were considered for the experiments: good, double-printed, and with an interrupted print. The images were randomly picked among the available ones for that particular class. To assess the features' relevance of a particular forecast, LIME[16] was used, considering the top 1, 3, 5, 7, and 13 ranked features.

Figure 4: GradCAM activation maps for ResNet-18 layers 1-4 and four layers combined.

The GradCAM images were generated for ResNet-18 layers 1-4 and another image considering the four layers. To understand where the underlying model focused, we created GradCAM activation maps contrasting the image against itself (see Fig. 4). The cosine similarity between the imputed vector and the image embedding was computed across test samples (880 samples: 679 good, 58 double-printed, and 143 related to interrupted printing). The mean similarity and standard deviation were used to assess whether the imputation strategy increased the similarity or contrast between the imputed vector and the image embedding.

The GradCAM images were generated by computing the cosine similarity between the image embedding and the feature vector generated considering three strategies described in Table 1. A sample of the resulting activation maps were visually assessed and are reported in Section 4.

The experiments were designed to understand which imputation strategy works the best. A detailed analysis regarding how top-ranked features affect the activation maps was omitted due to the brevity of the paper.

Strategy	Top-ranked feature	Selected on Feature Selection	Irrelevant
TOZ	True value	One	Zero
TZZ	True value	Zero	Zero
TRR	True value	Random	Random

Table 1: Value imputation strategies considering the image embedding, the features selected during the feature selection process, and the classifier's top-ranked features.

4 RESULTS

Imputation strategy	Image class	Layers			
		1	2	3	4
TOZ	Good	0.27±0.01	0.27±0.01	0.27±0.01	0.27±0.01
	Double-printed	0.31±0.02	0.31±0.02	0.31±0.02	0.31±0.02
	Interrupted print	0.27±0.01	0.27±0.01	0.27±0.01	0.27±0.01
TZZ	Good	0.21±0.04	0.21±0.04	0.21±0.04	0.21±0.04
	Double-printed	0.24±0.03	0.24±0.03	0.24±0.03	0.24±0.03
	Interrupted print	0.22±0.04	0.22±0.04	0.22±0.04	0.22±0.04
TRR	Good	0.46±0.02	0.46±0.02	0.46±0.02	0.46±0.02
	Double-printed	0.48±0.03	0.48±0.03	0.48±0.03	0.48±0.03
	Interrupted print	0.46±0.02	0.46±0.02	0.46±0.02	0.46±0.02

Table 2: Value imputation strategies considering the image embedding, the features selected during the feature selection process, and the classifier's top-ranked features.

As described in Table 1, three imputation strategies were considered. The cosine similarity computed between the vector created with the imputation strategy and the embedding (considering the top 13 features) is reported in Table 2. A higher similarity between the imputed vector and the image embedding means that a wider area of the activation map will be highlighted, blurring relevant information where the top features point to in the image. The less informative imputation strategy was TRR, which consistently showed high cosine similarity across layers for all defect types. On the other hand, TZZ achieved the best results regardless of the defect and layer considered. Imputing selected features with one had a detrimental effect, given it increased the similarity between the imputed vector and the embedding. Nevertheless, the similarity was usually between 0.10 and 0.20 points below that reported with the TRR imputation strategy.

For visual assessment, activation maps for different imputation strategies obtained for the top 13 features are displayed in Fig. 5. When comparing TZZ and TRR strategies, we found that for layer one, TZZ for the double-printed image focused on the top contour of characters, and for the interrupted print highlighted regions of relevance. In contrast, TRR did not highlight any region for the double-printed image and highlighted fewer regions for the interrupted print when compared to TZZ. For layer two, TZZ for the image of the non-defective product displayed some artifacts but included areas covering characters' contours, too. Furthermore, for the double-printed and interrupted print images, it covered relevant regions. TRR, on the other hand, highlighted different regions, which, for the good and double-printed images, were mostly irrelevant. For layer three, TZZ highlighted mostly irrelevant areas for the image of the non-defective product, except for the character "S". For the double-printed image, the beginning and end of the words are highlighted, while for the interrupted prints, the highlighted areas covered places where defects were observed. TRR, on the other hand, for the good image, covered two-thirds of the image, and for the double-printed, it highlighted most of the areas highlighted with the PZZ strategy. Nevertheless, for the interrupted print, most focus was placed on the lower part of the "P" char, while also two artifacts were encountered. Finally, for the fourth layer, TZZ has mostly focused on the upper word (Philips), while TRR's focus was mostly on the lower part of the image, still covering some relevant areas.

When comparing the TZZ and TOZ approaches, we found that for layer one, TOZ results in less strongly highlighted regions: most of the highlighted regions present in TZZ vanished, and just in the good image, a few spots appeared that were not present at the TZZ activation map. The original regions are highlighted for layer two, but new regions were included, mostly covering areas of interest. The highlighted areas for a double-printed image related to TZZ and TOZ activation maps were consistent for layer three. Nevertheless, TOZ highlighted different regions for the good and interrupted print images. The regions highlighted for the interrupted print image were irrelevant to defect detection. When considering the last layer, the highlighted areas were mostly the same for TZZ and TOZ. Nevertheless, an additional region was introduced in the good and interrupted print images, covering the lower text.

From the visual assessment described above, we conclude that activation maps obtained with the TZZ imputation method lead to the best explanations.

5 CONCLUSIONS

This work has researched how information regarding feature importance when using image embeddings can be used and propagated back to generate activation maps and highlight regions of the image considered relevant to a particular forecast. The proposed approach was evaluated on images of a real-world industrial use case. The similarity metrics and visual evaluation show that the best value imputation strategy is TZZ, which considers assigning the actual embedding value to relevant features

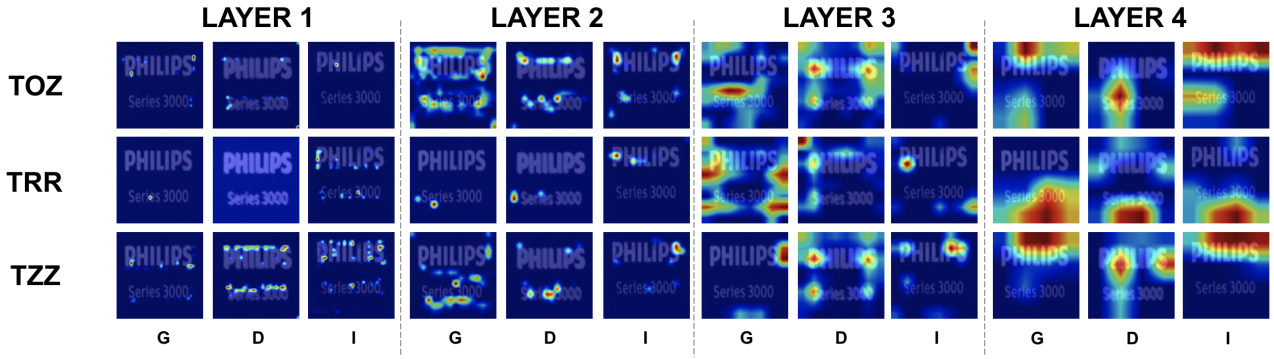

Figure 5: GradCAM activation maps for ResNet-18 layers 1-4 considering only the top 13 features for this particular forecast and three imputation strategies (TOZ, TZZ, and TRR) for three image types (good (G), double-printed (D), and interrupted prints (I)).

and masking the rest of the embedding with zeroes. Nevertheless, it must be emphasized that a broader set of experiments must be considered to generalize these conclusions. While this research only considered local explanations, the feature relevance could be considered at a global level, and the same approach was leveraged to visualize their influence on a particular image. Future work will focus on a more comprehensive evaluation of the proposed methodology to understand how it performs, how the number of selected features influences the activation maps and possible shortcomings.

ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency and the European Union's Horizon 2020 program project STAR under grant agreement number H2020-956573.

REFERENCES

[1] Imran Ahmed, Gwanggil Jeon, and Francesco Piccialli. 2022. From artificial intelligence to explainable artificial intelligence in industry 4.0: a survey on what, how, and where. *IEEE Transactions on Industrial Informatics* 18, 8 (2022), 5031–5042.

[2] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion* 58 (2020), 82–115.

[3] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. 2018. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 839–847.

[4] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. 2019. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems* 32 (2019).

[5] Arun Das and Paul Rad. 2020. Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371* (2020).

[6] Rachel Lea Draelos and Lawrence Carin. 2020. Use HiResCAM instead of Grad-CAM for faithful explanations of convolutional neural networks. *arXiv preprint arXiv:2011.08891* (2020).

[7] Gautam Dutta, Ravinder Kumar, Rahul Sindhwani, and Rajesh Kr Singh. 2021. Digitalization priorities of quality control processes for SMEs: A conceptual study in perspective of Industry 4.0 adoption. *Journal of Intelligent Manufacturing* 32, 6 (2021), 1679–1698.

[8] Ruth C Fong and Andrea Vedaldi. 2017. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE international conference on computer vision*. 3429–3437.

[9] Ruigang Fu, Qingyong Hu, Xiaohu Dong, Yulan Guo, Yinghui Gao, and Biao Li. 2020. Axiom-based grad-cam: Towards accurate visualization and explanation of cnns. *arXiv preprint arXiv:2008.02312* (2020).

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[11] Jianping Hua, Zixiang Xiong, James Lowey, Edward Suh, and Edward R Dougherty. 2005. Optimal number of features as a function of sample size for various classification rules. *Bioinformatics* 21, 8 (2005), 1509–1515.

[12] Tambiama André Madiega. 2021. Artificial intelligence act. *European Parliament: European Parliamentary Research Service* (2021).

[13] Dang Minh, H Xiang Wang, Y Fen Li, and Tan N Nguyen. 2022. Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review* (2022), 1–66.

[14] Sajid Nazir, Diane M Dickson, and Muhammad Usman Akram. 2023. Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks. *Computers in Biology and Medicine* (2023), 106668.

[15] Cecilia Panigutti, Ronan Hamon, Isabelle Hupont, David Fernandez Llorca, Delia Fano Yela, Henrik Junklewitz, Salvatore Scalzo, Gabriele Mazzini, Ignacio Sanchez, Josep Soler Garrido, et al. 2023. The role of explainable AI in the context of the AI Act. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 1139–1150.

[16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[17] Gesina Schwalbe and Bettina Finzel. 2023. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery* (2023), 1–59.

[18] A Carlisle Scott, William J Clancey, Randall Davis, and Edward H Shortliffe. 1977. *Explanation capabilities of production-based consultation systems.* Technical Report. STANFORD UNIV CA DEPT OF COMPUTER SCIENCE.

[19] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*. 618–626.

[20] Bas HM Van der Velden, Hugo J Kuijf, Kenneth GA Gilhuijs, and Max A Viergever. 2022. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Medical Image Analysis* 79 (2022), 102470.

[21] Giulia Vilone and Luca Longo. 2021. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion* 76 (2021), 89–106.

[22] Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu. 2019. Explainable AI: A brief survey on history, research areas, approaches and challenges. In *CCF international conference on natural language processing and Chinese computing*. Springer, 563–574.

[23] Li Da Xu, Eric L Xu, and Ling Li. 2018. Industry 4.0: state of the art and future trends. *International journal of production research* 56, 8 (2018), 2941–2962.

[24] Vitjan Zavrtanik, Matej Kristan, and Danijel Skočaj. 2021. Draem-a discriminatively trained reconstruction embedding for surface anomaly detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8330–8339.