Towards Complementary Explanations Using Deep Neural Networks

Wilson Silva[1,2(✉)], Kelwin Fernandes[1,2], Maria J. Cardoso[2,3,4],
and Jaime S. Cardoso[1,2]

[1] Faculdade de Engenharia, Universidade do Porto, Porto, Portugal
[2] INESC TEC, Porto, Portugal
wilson.j.silva@inesctec.pt
[3] Faculdade de Ciências Médicas, Universidade NOVA de Lisboa, Lisbon, Portugal
[4] Breast Unit, Champalimaud Foundation, Lisbon, Portugal

Abstract. Interpretability is a fundamental property for the acceptance of machine learning models in highly regulated areas. Recently, deep neural networks gained the attention of the scientific community due to their high accuracy in vast classification problems. However, they are still seen as black-box models where it is hard to understand the reasons for the labels that they generate. This paper proposes a deep model with monotonic constraints that generates complementary explanations for its decisions both in terms of style and depth. Furthermore, an objective framework for the evaluation of the explanations is presented. Our method is tested on two biomedical datasets and demonstrates an improvement in relation to traditional models in terms of quality of the explanations generated.

Keywords: Interpretable machine learning · Deep neural networks
Explanations · Aesthetics evaluation · Dermoscopy

1 Introduction

In the most recent years many machine learning models are replacing or helping humans in decision-making scenarios. The recent success of deep neural networks (DNN) in the most diverse applications led to a widespread use of this technique. Nonetheless, their high accuracy is not accompanied by high interpretability. On the contrary, they remain mostly as black-box models. In this way and despite the success of DNN, in areas such as medicine and finance, which have legal and safety constraints, their use is somehow restricted. Therefore, and in order to take advantage of the DNN potential, it is critical to develop robust strategies to explain the behavior of the model. In the literature it is possible to find several different strategies to generate reasonable and perceptible explanations for machine learning model's behavior. However, those strategies can be grouped in three clusters of interpretable methods: pre-model, in-model and post-model [6].

One of the options is to consider the relevance of example-based explanations in human reasoning to try to make sense about the data we are dealing with. The

main idea here is that a complex data distribution might be easily interpretable considering prototypical examples. Considering that the goal is to understand the data before building any machine learning model, one can consider this strategy as interpretability before the model, i.e. pre-model.

An alternative is to build interpretability in the model itself. Inside this group, models can be based in rules, cases, sparsity and/or monotonicity. Rule-based models are characterized by a set of rules which describe the classes and define predictions. One problem typically related with this strategy is the size of the interpretable model. In order to solve this issue, Wang *et al.* [9] proposed a Bayesian framework to control the size and shape of the model. Nevertheless, a rule-based model is as interpretable as its original features are. Leveraging once more the power of examples in human understanding but now with the aim of building a machine learning model, case-based methods appear as serious competitors in the explainability challenge. In [5], the authors present a model that generates its explanations based on cluster divisions. Each cluster is characterized by a prototype and a set of defining features. From this, it is possible to deduce that the model's explanations are limited by the quality of the prototype. Sparsity is also an important property to achieve interpretability. With a limited number of activations it is easier to determine what were the events that determined the model's decision. However, if the decision can not be made with just a few activations, sparsity can decisively affect the accuracy of the model. Another way of facilitating the model interpretability is to guarantee the monotonicity of the learnt function in relation to some of the inputs [4].

Finally, interpretability can be performed after building a model. One of the options is sensitivity analysis, which consists on disturbing the input of the model and observing what happens to its output. In a computer vision context this could mean occlusions of some parts of the image [3]. One issue with sensitivity analysis is that a change in the input may not represent a realistic scenario in the data distribution. Other possibility is to create a new model capable of imitating the one which is giving the classification predictions. For instance, one can mimic a DNN with a more shallow [1] and, consequently, more interpretable network. However, it is not always the case that a simpler model exists. Lastly, we have interpretability given by investigation on hidden layers of deep convolutional neural networks [10].

### 1.1   Satisfying the Curiosity of Decision Makers

Human beings have different ways of thinking and learning [8]. There are people for whom a visual explanation is more easily apprehended and, on the contrary, there are people who prefer a verbal explanation. In order to satisfy all the decision makers, an interpretable model should be able to provide different styles of explanations and with different levels of granularity. Furthermore, it should present as many explanations as the decision maker needs to be confident about his/her decisions. It is also important to mention that some observations require more complex explanations than others, which reinforces the idea of different depth in the explanations.

2   Complementary Explanations Using Deep Neural Networks

In addition to their high accuracy in various classification problems, DNN have the ability to jointly integrate different strategies of interpretability, such as, the previously mentioned, case-based, monotonicity and sensitivity analysis. Thus, it is a model that presents itself at the forefront to satisfy the decision makers in their search for valuable and diverse explanations.
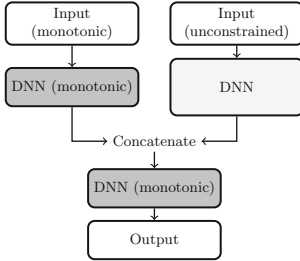
Fig. 1. Proposed DNN architecture.
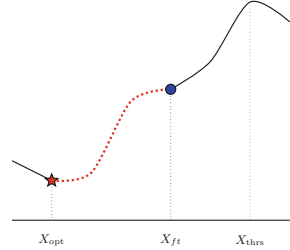
Fig. 2. Feature impact analysis.

We will focus on binary classification settings with a known subset of monotonic features. Without loss of generality, we will assume that monotonic features increase with the probability of observing the positive class. The proposed architecture consists on two independent streams of densely connected layers that process the monotonic and non-monotonic features respectively. We impose constraints on the weights of the monotonic stream to be positive to facilitate interpretability. Then, both streams are merged and processed by a sequence of densely connected layers with positive constraints. Thus, we are promoting that the non-monotonic stream maps its feature space into a latent monotonic space. It is expected that the non-monotonic features will require additional expressiveness to transform a non-monotonic space into a monotonic one. In this sense, we validate topologies where the non-monotonic stream has at least as many –and possibly more– layers than the monotonic stream. Figure 1 illustrates the proposed architecture.

Explanation by Local Contribution. To measure the contribution, C_{ft}, of a feature ft on the prediction y, we can find the assignment X_{opt} that approximates X to an adversarial example (see (1)):

$$(\bar{y} - f(X))^2 \qquad (1)$$

where $\bar{y} = 1 - y$ is the opponent class, $y \in \{0, 1\}$, and $f(X)$ is the estimated probability. We can use backpropagation with respect to ft to find the value X_{opt} (see Fig. 2) that minimizes (1). It is relevant to note that for monotonic features,

such value is known a priori. Since some features may have a generalized higher contribution than others, resulting in repetitive explanations, we balanced the contribution on the target variable with the range of the feature domain traversed from the initial value to the local minimum X_{opt}. Namely:

$$C_{ft} = |f(X) - f(X')| \cdot \frac{X_{ft} - X_{opt}}{X_{\max} - X_{\min}} \tag{2}$$

where X' is the input vector after assigning X_{opt} to the feature ft. Thus, the contribution can be measured by approximating X to the adversarial space. On the other hand, the inductive rule constructed for ft covers the space between X_{ft} and the value X_{thrs} where the probability of the predicted class is maximum.
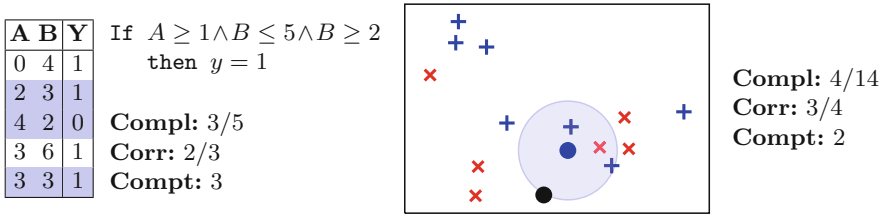
Explanation by Similar Examples. DNN are able to learn intermediate semantic representations adapted to the predictive task. Thus, we can use the nearest neighbors in the semantic space as an explanation for the decision. While the latent space is not fully interpretable, we can evaluate which features (and at which degree) impact the distance between two observations using sensitivity analysis. In this sense, two types of explanations can be extracted:

- **Similar:** the nearest neighbor in the latent space and what features make them similar.
- **Opponent:** the nearest neighbor from the opponent class in the latent space and what features make them different.

3   The Three Cs of Interpretability

Interpretability and explainability are tied concepts often used interchangeably. In this work, we will focus on local explanations of the predicted class, where individual explanations are provided for each observation. Despite the vast amount of effort that has been invested around interpretable models, the concept itself is still vaguely defined and lacks of a unified formal framework to assess it. The efficacy of an explanation depends on its ability to convince the target audience. Thus, it is surrounded by external intangible factors such as the background of the audience and its willingness to accept the explanation as a truth. While it is hard to fully assess the quality of an explanation, some proxy functions can be used to summarize the quality of a prediction under certain assumptions. Let us define an explanation as a simple model that can be applied to a local context of the data. A good explanation should maximize the following properties:

- **Completeness:** It should be susceptible of being applied in other cases where the audience can verify the validity of that explanation. e.g., the blue rows in Fig. 3 where the decision rule precondition holds and the observations within the same distance of the neighbor explanation (Fig. 3).
- **Correctness:** It should generate trust (i.e., be accurate). e.g., the label agreement between the blue rows and between the points inside the n-sphere.
- **Compactness:** It should be succinct. e.g., the number of conditions in the decision rule and the feature dimensionality of a neighbor-based explanation.

A	**B**	**Y**
0	4	1
2	3	1
4	2	0
3	6	1
3	3	1

If $A \geq 1 \wedge B \leq 5 \wedge B \geq 2$
  then $y = 1$

Compl: 3/5
Corr: 2/3
Compt: 3

Compl: 4/14
Corr: 3/4
Compt: 2

Fig. 3. Illustration of explanation quality for decision rules and KNN (where the black dot is the new observation and the blue dot is the nearest-neighbor).

4   Experimental Assessment

We validate the performance of the proposed methodology on two applications. First, we consider the post-surgical aesthetic evaluation (i.e., poor, fair, good, and excellent) of breast cancer patients [2]. The dataset has 143 images with 23 high-level features describing breast asymmetry in terms of shape, global and local color (i.e., scars). The second application consists on the classification of dermoscopy images in three classes: common nevus, atypical nevus and melanoma. The dataset [7] has 14 features from 200 patients describing the presence of certain colors on the nevus and abnormal patterns such as asymmetry, dots, streaks, among others. In both cases, we consider binary discretizations of the problem (see Table 1). In this work, we assume features are already extracted in a previous stage of the pipeline. However, the entire pipeline covering feature extraction and model fitting could be learned end-to-end using intermediate supervision on the feature representation.
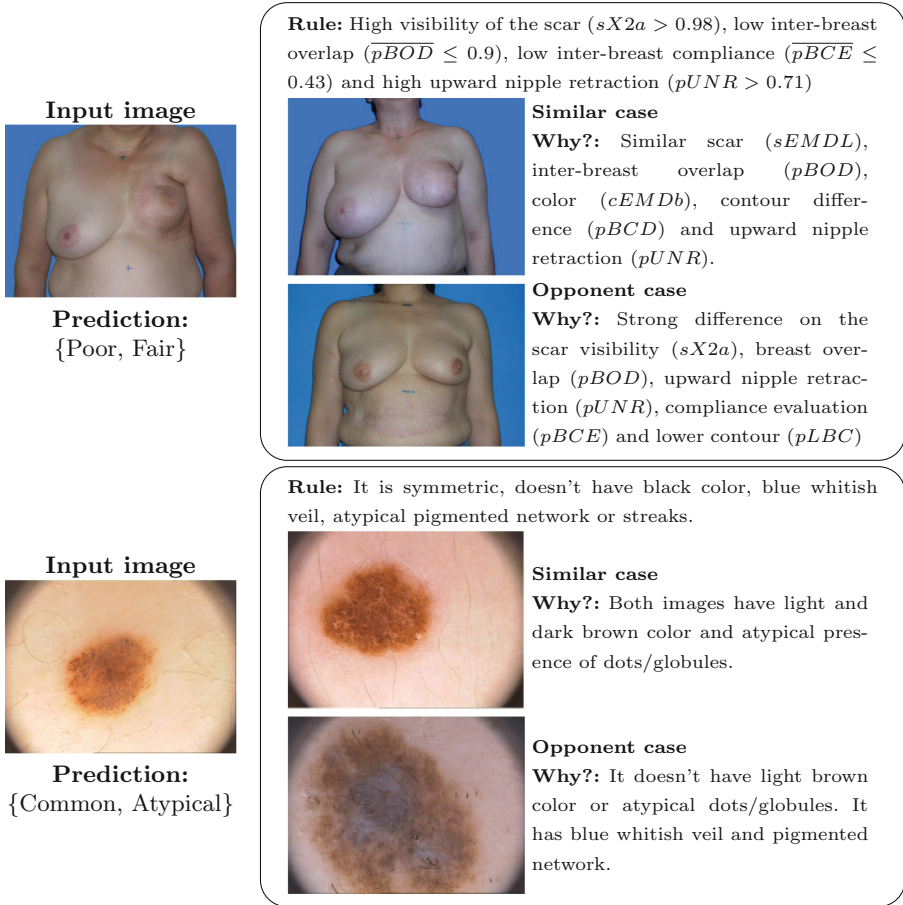
We compare the performance of the proposed DNN against classical interpretable models: a decision tree (DT) with bounded depth learned with the CART algorithm, and a Nearest Neighbor classifier (KNN with K = 1). We used stratified 10-fold cross-validation to choose the best hyper-parameter configuration and to generate the explanations. We explore DNN topologies with depth between 1 and 3 per block (see Fig. 1). We show in Table 1 the model performance of the three models. DNN achieved better performance than the remaining classifiers in most cases.

To measure the quality of the explanations we used accuracy for correctness, the fraction of the training set covered by the explanation as completeness, and the size in bytes of the explanation (the lower the better) after compression using the standard Deflate algorithm. Despite this compactness metric doesn't reflect the actual complexity of the explanations, it is a proxy function to define it under the assumption that the time to understand an explanation is proportional to its length. We generate explanations that account for 95% of the feature impact and embedding distance. This value can be adapted to produce more general/global or customized/local explanations. As can be seen in the

Table 1. Quality of the predictions in terms of area under the ROC and Precision-Recall curves. Quality of the explanations in terms of correctness (Corr), completeness (Compl), and compactness (Compt).

Binarization	Model	Predictions		Explanations			
		ROC	PR	Type	Corr	Compl	Compt
BCCT [2]: Breast aesthetics							
Excellent *vs.* Good, Fair, Poor	DT	71.96	92.19	Rule	75.52	3.82	31.97
	1-NN	67.37	90.74	Similar	89.27	3.25	95.94
				Opponent	72.96	80.84	96.00
	DNN	**80.61**	96.55	Similar	85.69	95.20	124.94
				Opponent	92.04	46.87	149.68
				Rule	99.91	3.69	62.59
Excellent, Good *vs.* Fair, Poor	DT	85.18	75.20	Rule	51.75	3.16	30.00
	1-NN	52.81	39.49	Similar	85.69	2.98	95.94
				Opponent	54.76	91.26	95.97
	DNN	**86.78**	**82.82**	Similar	72.52	17.34	80.36
				Opponent	81.16	31.28	138.00
				Rule	98.89	2.33	48.59
Excellent, Good Fair *vs.* Poor	DT	**94.20**	**74.92**	Rule	76.92	6.71	17.0769
	1-NN	54.42	20.63	Similar	94.45	3.01	95.94
				Opponent	84.42	85.33	96.00
	DNN	91.03	73.00	Similar	87.25	1.46	79.79
				Opponent	92.82	67.86	157.81
				Rule	99.88	5.48	58.44
PH^2 [7]: Dermoscopy images							
Common *vs.* Atypical, Melanoma	DT	97.60	97.90	Rule	43.00	5.03	13.10
	1-NN	94.37	94.29	Similar	94.97	5.56	15.29
				Opponent	59.42	81.38	15.94
	DNN	**99.74**	**99.83**	Similar	97.11	39.00	19.32
				Opponent	74.59	70.61	37.69
				Rule	98.86	38.83	16.27
Common, Atypical *vs.* Melanoma	DT	95.55	81.63	Rule	82.00	5.82	19.00
	1-NN	80.94	63.67	Similar	94.81	5.70	15.23
				Opponent	69.75	86.98	21.25
	DNN	**96.02**	**89.30**	Similar	91.49	8.15	33.27
				Opponent	84.02	62.12	46.24
				Rule	97.89	44.84	23.65

results, the proposed model is able to achieve the best performance in correctness results for rule explanations. For case-based explanation, the 1-NN approach with similar prototype achieves better performance in some cases at the expense of completeness. Therefore, we validate that besides having a good predictive performance in terms of classification, we can use DNN to produce explanations with high quality. Figure 4 shows some explanations produced by the DNN for both datasets.

Input image

Prediction:
{Poor, Fair}

Rule: High visibility of the scar ($sX2a > 0.98$), low inter-breast overlap ($\overline{pBOD} \leq 0.9$), low inter-breast compliance ($\overline{pBCE} \leq 0.43$) and high upward nipple retraction ($pUNR > 0.71$)

Similar case
Why?: Similar scar ($sEMDL$), inter-breast overlap ($pBOD$), color ($cEMDb$), contour difference ($pBCD$) and upward nipple retraction ($pUNR$).

Opponent case
Why?: Strong difference on the scar visibility ($sX2a$), breast overlap ($pBOD$), upward nipple retraction ($pUNR$), compliance evaluation ($pBCE$) and lower contour ($pLBC$)

Input image

Prediction:
{Common, Atypical}

Rule: It is symmetric, doesn't have black color, blue whitish veil, atypical pigmented network or streaks.

Similar case
Why?: Both images have light and dark brown color and atypical presence of dots/globules.

Opponent case
Why?: It doesn't have light brown color or atypical dots/globules. It has blue whitish veil and pigmented network.

Fig. 4. Visualization of the explanations. In the BCCT dataset we are considering the binary classification problem: {Poor, Fair} *vs.* {Good, Excellent}. Regarding the PH^2, the classification problem comes down to {Common, Atypical} *vs.* {Melanoma}. $\overline{pBOD}$ and $\overline{pBCE}$ represent the negation of the original features, $pBOD$ and $pBCE$, and are presented to make the explanation more intuitive.

5   Conclusion

In order for a machine learning model to be adopted in highly regulated areas such as medicine and finance, it needs to be interpretable. However, interpretability is a vague concept and lacks an objective framework for evaluation.

In this work, we proposed a DNN model able to generate complementary explanations both in terms of type and granularity. Moreover, there can be as many explanations as the ones the decision maker considers necessary to satisfy his/her doubts. We also define some proxy functions that summarize relevant

aspects of interpretability, namely, completeness, correctness and compactness. This way we get an objective framework to evaluate the explanations generated.

The model is evaluated in two biomedical applications: post-surgical aesthetic evaluation of breast cancer patients and classification of dermoscopy images. Both the quantitative and qualitative results of our model show an improvement in the quality of the explanations generated compared to other interpretable models. Future work will focus on extending this model to ordinal and multiclass classification.

References

1. Ba, J., Caruana, R.: Do deep nets really need to be deep? In: Advances in Neural Information Processing Systems 27, pp. 2654–2662. Curran Associates, Inc. (2014)
2. Cardoso, J.S., Cardoso, M.J.: Towards an intelligent medical system for the aesthetic evaluation of breast cancer conservative treatment. Artif. Intell. Med. **40**, 115–126 (2007)
3. Fernandes, K., Cardoso, J.S., Astrup, B.: A deep learning approach for the forensic evaluation of sexual assault. Pattern Anal. Appl. **21**, 629–640 (2018)
4. Gupta, M., et al.: Monotonic calibrated interpolated look-up tables. J. Mach. Learn. Res. **17**(109), 1–47 (2016)
5. Kim, B., Rudin, C., Shah, J.A.: The Bayesian case model: a generative approach for case-based reasoning and prototype classification. In: Advances in Neural Information Processing Systems 27, pp. 1952–1960. Curran Associates, Inc. (2014)
6. Kim, B., Doshi-Velez, F.: Interpretable machine learning: the fuss, the concrete and the questions. In: ICML Tutorial on Interpretable Machine Learning (2017)
7. Mendonça, T., Ferreira, P.M., Marques, J.S., Marcal, A.R.S., Rozeira, J.: PH2 - a dermoscopic image database for research and benchmarking. In: 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 5437–5440, July 2013
8. Pashler, H., McDaniel, M., Rohrer, D., Bjork, R.: Learning styles: concepts and evidence. Psychol. Sci. Public Interes. **9**(3), 105–119 (2008)
9. Wang, T., Rudin, C., Doshi-Velez, F., Liu, Y., Klampfl, E., MacNeille, P.: A bayesian framework for learning rule sets for interpretable classification. J. Mach. Learn. Res. **18**(70), 1–37 (2017)
10. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 818–833. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10590-1_53