Contrastive Explanations to Classification Systems Using Sparse Dictionaries

A. Apicella, F. Isgrò(✉), R. Prevete, and G. Tamburrini

Dipartimento di Ingegneria Elettrica e delle Teconologie dell'Informazione,
Università degli Studi di Napoli Federico II, Naples, Italy
{andrea.apicella,francesco.isgro}@unina.it

Abstract. Providing algorithmic explanations for the decisions of machine learning systems to end users, data protection officers, and other stakeholders in the design, production, commercialisation and use of machine learning systems pipeline is an important and challenging research problem. Much work in this area focuses on image classification, where the required explanations can be given in terms of images, therefore making explanations relatively easy to communicate to end-users. For a classification problem, a contrastive explanation tries to understand why the classifier has not answered a particular class, say B, instead of the returned class A. Sparse dictionaries have been recently used to identify local image properties as main ingredients for a system producing humanly understandable explanations for the decisions of a classifier developed based on machine learning methods. In this paper, we show how the system mentioned above can be extended to produce contrastive explanations.

Keywords: XAI · Explainable artificial intelligence · Machine learning · Sparse coding · Contrastive explanations

1 Introduction

Machine Learning (ML) techniques make possible to develop systems that learn from observations. Many ML techniques (e.g., Support Vector Machines (SVM) and Deep Neural Networks (DNN)) give rise to systems the behaviour of which is often hard to interpret [18]. A crucial ML interpretability issue concerns the generation of explanations for an ML system behaviour that are understandable to a human being. In general, this issue is addressed as a scientific and technological problem by so-called explainable artificial intelligence (XAI) [1,9,20,23]. Providing XAI solutions to the ML explainability problem is important for many AI and computer science research areas: to improve intelligent systems design,

The research presented in this paper was partially supported by the national project Perception, Performativity and Cognitive Sciences (PRIN Bando 2015, cod. 2015TM24JS_009).

E. Ricci et al. (Eds.): ICIAP 2019, LNCS 11751, pp. 207–218, 2019.
https://doi.org/10.1007/978-3-030-30642-7_19

testing and revision processes, to make the rationale of automatic decisions more transparent to end users and systems managers, thereby leading to better forms of HCI and HRI involving learning systems, to improve interactions between learning agents in Distributed AI, and so on. Providing a solution to the ML explainability problem is also important from an ethical and legal viewpoint. ML systems are being increasingly used to make or to support decisions that have an impact on the life of persons, including career development, court decisions, medical diagnosis, insurance risk profiles and loan decisions.

Various senses of interpretability and explainability for learning systems have been identified and analysed [9], and various approaches to overcoming their opaqueness are now being pursued [11,27]. For example, in [24] a series of techniques for the interpretation of DNN are discussed, and in [20] a wide variety of motivations underlying interpretability needs are examined, thereby refining the notion of interpretability in ML systems. In the context of this multifaceted interpretability problem [34,35], we focus on the issue of what it is to explain the behaviour of ML perceptual classification systems for which only I/O relationships are accessible, i.e., the learning system is seen as a black-box. In literature, this type of approach is known as *model agnostic* [31].

Various model agnostic approaches have been proposed to give *global* explanations by exhibiting a class prototype to which the input data can be associated [11,24,27,34]. These explanations are given in response to requests usually expressed as why-questions: “Why was input x associated to class C?”. Specific why-questions which may arise in connection with actual learning systems are: “Why was this loan application rejected?” and “Why was this image classified as a fox?”. However, prototypes often make rather poor explanations available. For instance, if an image x is classified as “fox”, the explanation provided by means of a fox-prototype is nothing more than a “because it looks like this” explanation: one would not be put in the position to understand what features (parts) of the prototype are associated to what characteristics (parts) of x. In order to go beyond this level of understanding, instead of merely giving the user a global explanation, one might attempt to provide a *local* explanation, which highlights salient parts of the input [31]. Furthermore, [13,23] highlight that an human explanation of an event is often given in *contrastive* terms, that is, instead of trying to answer to the question “why this outcome?”, a possible answer to the question “why this outcome and not another one?” is given. This result can be reached considering, during the generation of the explanation, an event that did not occur instead of the event that really happened, for example searching for an explanation on the reasons behind an classifier returns “dog” as answer to a given input image and not “cat”. So, in contrastive explanation approaches, a different hypothetical outcome, which [19] calls the “foil”, is always used to build the explanation.

In this paper, we exploit a model agnostic framework that returns local explanations of classifications [2,29] in order to obtain an explanation in contrastive terms. This framework, which is based on *dictionaries* of local and humanly interpretable elements of the input, can be functionally described as a three entities model, composed of an *Oracle* (an ML system, e.g. a classifier), an *Interrogator* raising explanations requests about the Oracle's responses, and a *Mediator* helping the Interrogator to understand the answer given by the Oracle. In this framework, local explanations are provided by a module (the Mediator) which is different from the classifier itself. The Mediator plays the crucial explanatory role, by advancing hypotheses on what humanly interpretable elements are likely to have influenced the Oracle output, building explanations both in classical terms ("why P?") and in contrastive terms ("why P and not Q?"). More specifically, elements are computed which represent humanly interpretable features of the input data, with the constraint that both prototypes and input can be reconstructed as linear combinations of these elements. Thus, one can establish meaningful associations between key features of the prototype and key features of the input. To this end, we exploit the representational power of sparse dictionaries learned from the data, where atoms of the dictionary selectively play the role of humanly interpretable elements, insofar as they afford a local representation of the data. Indeed, these techniques provide data representations that are often found to be accessible to human interpretation [22]. The dictionaries are obtained by a Non-negative Matrix Factorisation (NMF) method [4,14,17], and the explanations are determined using an Activation-Maximisation (AM) [11,34] based technique.

The paper is organised as follows: Sect. 2 briefly reviews related approaches, in Sect. 3 we present the overall architecture; experiments and results are discussed in Sect. 4, while Sect. 5 is devoted to concluding remarks and future developments.

2 Related Work

In recent years, various attempts have been made to interpret and explain the output of a classification system. Initial attempts concerned SVM classifiers (see for example [28]) or rule-based systems [6,8].

In the neural network context, recent surveys on explainable AI are proposed in [1,12,30,40]. A significant attempt to explain in terms of images what a computational neural unit computes is found in [11] using the *Activation Maximisation* method. AM-like approaches applied to CNN were proposed in [21,34]. Additional attempts to give interpretability to CNNs were proposed in [37] and [10], where Deconvolutional Network (already presented by [38] as a way to do unsupervised learning) and *up-convolutional network* are proposed, while [26,27] uses an image generator network (similar to GANs) as priors for AM algorithm to produce synthetic preferred images. In these approaches, explanations are given in terms of prototypes or approximate input reconstructions. However, one does not take into account the issue whether the given explanations are in some manner interpretable by humans. Moreover, the proposed approaches seem to be model-specific for CNN, differently from our model which is to be considered as model-agnostic, and consequently applicable in principle to any classifier. From another point of view, [36] studies the influence on the output of hardly perceptible perturbation on the input, empirically showing that it is possible to arbitrarily change the network's prediction even when the input is left apparently unchanged. Although this type of noise is extremely unlikely to occur in realistic situations, the fact that such noise is imperceptible to an observer opens interesting questions about the semantics of network components. However, approaches of this kind are quite distant from our present concerns, insofar as they focus on entities that are hardly meaningful to humans. Important works are also made into [3,5,25] where Pixel-Wise Decomposition, Layer-Wise Relevance propagation ad Deep Taylor Decomposition are presented. [33] builds explanations as difference in output from a "reference" output in terms of the difference of the input from a "reference" input.

[41] presents a work based on *prediction difference analysis* [32] where a features relevance vector is built which estimates how much each feature is "important" for the classifier to return the predicted class. In [31] , the model-agnostic explainer LIME is proposed, which takes into account the model behaviour in the proximity of the instance being predicted. The LIME framework is more similar to our approach than the other approaches mentioned in this section, and many other approaches found in the literature. The LIME framework differs from our own mainly in its use of super-pixels instead of a learned dictionary constrained in order to have a compact representation.

In [39] a XAI methods based on the contrastive explanations is proposed. However, this method relies on Deep Neural Network (specifically a CNN), making this approach model-specific, differently from our proposed model which is model-agnostic, that is independent by the chosen model to explain.

3 Proposed Approach

Given an oracle Ω, an input $\boldsymbol{x}$ and an Ω's answer $\hat{c}$ (regardless of whether it is correct or not), we want to give an explanation of the answer provided by the model Ω that is humanly interpretable. As we want to obtain humanly interpretable elements which, combined together, can provide an acceptable explanation for the choice made by Ω, we search for an explanation having the following qualitative properties:

1. the explanation must be expressed in terms of a *dictionary* V whose elements (atoms) are easily understandable by an interrogator;
2. the elements of the dictionary V have to represent "local properties" of the input $\boldsymbol{x}$;
3. the explanation must be composed by few dictionary elements.

We claim that considering as elements atoms of a sparse coding from a sparse dictionary, and using sparse coding methods together with an AM-like algorithm we obtain explanations satisfying the properties described above. Furthermore, since the proposed method gives explanations in terms of relevant components (atoms) which contributed to the classifier decision, we take advantage of this property to generate discriminative explanations comparing the explanation produced for the real classifier outcome with the explanation produced for a contrast class given the same input. We think that, showing explanations generated for different classes can help in understanding the reason behind the "preference" given by an Oracle to an answer instead of another one.

3.1 Sparse Dictionary Learning

The first step of the proposed approach consists in finding a "good" dictionary V that can represent data in terms of humanly interpretable atoms.

Let us assume that we have a set $D = \{(\boldsymbol{x}^{(1)}, c^{(1)}), (\boldsymbol{x}^{(2)}, c^{(2)}). \dots, (\boldsymbol{x}^{(n)}, c^{(n)})\}$ where each $\boldsymbol{x}^{(i)} \in \mathbb{R}^d$ is a column vector representing a data point, and $c^{(i)} \in C$ its class. We can learn a Dictionary $V \in \mathbb{R}^{d \times k}$ of k atoms across multiple classes and an encoding $H \in \mathbb{R}^{k \times n}$ s.t. $X = VH + \epsilon$ where $X = (\boldsymbol{x}^{(1)}|\boldsymbol{x}^{(2)}|\dots|\boldsymbol{x}^{(n)})$ and ϵ is the error introduced by the coding. Every column $\boldsymbol{x}^{(i)}$ in X can be expressed as $\boldsymbol{x}^{(i)} = V\boldsymbol{h}_i$ with h_i $i-$th column of H. The dictionary forms the basis of our explanation framework for an ML system.

We selected as dictionary learning algorithm an NMF scheme [17] with the additional sparseness constraint proposed by [14]; this choice is motivated by the fact that it respects our requirements described above, giving a "local" representation of data, and *non-negativity*, that ensures only additive operations in data representations, giving a better human understanding with respect to other techniques. The sparsity level can be set using two parameters γ_1 and γ_2 which control the sparsity on the dictionary and the encoding, respectively.

3.2 Explanation Maximisation

Unlike traditional dictionary-based coding approaches, our main goal is not to get an "accurate" representation of the input data, but to get a representation that helps humans to understand the decision taken by a trained model. To this aim, we modify the AM algorithm so that, instead of looking for the input that just maximises the answer of the model, it searches for the dictionary-based encoding $\boldsymbol{h}$ that maximises the answer and, at the same time, is sparse enough but without being "too far" from the original input $\boldsymbol{x}$. More formally, indicating with $\Pr(\hat{c}|\boldsymbol{x})$ the probability given by a learned model that input $\boldsymbol{x}$ belongs to class $\hat{c} \in C$, V the chosen dictionary, $S(\cdot)$ a sparsity measure, the objective function that we optimise is

$$\max_{\boldsymbol{h} \geq 0} \log \Pr\big(\hat{c}|V\boldsymbol{h}\big) - \lambda_1 ||V\boldsymbol{h} - \boldsymbol{x}||_2 + \lambda_2 S\big(\boldsymbol{h}\big) \tag{1}$$

where λ_1, λ_2 are hyper-parameters regulating the input reconstruction and the encoding sparsity level, respectively. The first regularisation term leads the algorithm to choose dictionary atoms that, with an appropriate encoding, form a good representation of the input, while the second regularisation term ensures a certain sparsity degree, i.e., that only few atoms are used. The $\boldsymbol{h} \geq 0$ constraint ensures that one has a purely additive encoding. Thus, each h_i, $\forall i. 1 \leq i \leq d$,

```
Algorithm 1: Explanation Maximization procedure
Input: data point $x \in \mathbb{R}^d$, the output class $\hat{c}$ ,learned model $\Gamma$, a dictionary
       $V \in \mathbb{R}^{d \times k}$, $\lambda_1, \lambda_2$
Output: the encoding $h \in \mathbb{R}^d$
h ∼ $U^d(0,1)$;
while ¬ converge do
    $r \leftarrow Vh$;
    $h \leftarrow \arg\max_{h} \Pr\big(\hat{c}|r;\Gamma\big) - \lambda_1||r - x||_2$;
    $h \leftarrow \text{proj}(h, \lambda_2)$;                ▷ proj(·,·) is given by [14]
end
return h ;
```

measures the "importance" of the i-th atom. Equation 1 is solved by using a standard gradient ascent technique, together with a projection operator given by [14] that ensures both sparsity and non-negativity. The complete procedure is reported in Algorithm 1.

3.3 Contrastive Explanation Maximisation

The aim of this we paper is to obtain a contrastive explanation approach exploiting the EM procedure described in Sect. 3.2. We remember that, instead of answering to the question "why the classifier returns the class P?", contrastive explanations wants to answer to the question "why the Oracle returns the class P and not the class Q?". The described EM procedure generates a possible explanation searching for a good subset of atoms which pushes the classifier toward the predicted class and, at the same time, is similar enough to the input under investigation. We can easily use the same procedure to push the classifier towards a contrastive class, so searching for a good set of atoms which is again near enough to the input but that gives a different outcome if fed to the classifier. An answer to the question "why the Oracle returns the class P and not the class Q?" can be given inspecting the difference between atoms in the generated explanations. For example, in a dataset of letters, if I have an image of an "e" and a classifier gives the correct class, I expect that the explanation of "why is it an "e"? " differs from the explanation of, for example, "why should it be a "c"?" by the use of some atom representing a centre line which characterises the "e" letter respect to the "c" letter. In other words, we search for two (or more) good enconding h_{c^*} and $h_{\overline{c}}$ such that

$$\begin{aligned} \boldsymbol{h}_{c^*} &= \arg\max_{\boldsymbol{h}\geq 0} \log \Pr\big(c^*|V\boldsymbol{h}\big) - \lambda_1||V\boldsymbol{h} - \boldsymbol{x}||_2 + \lambda_2 S\big(\boldsymbol{h}\big) \\ \boldsymbol{h}_{\overline{c}} &= \arg\max_{\boldsymbol{h}\geq 0} \log \Pr\big(\overline{c}|V\boldsymbol{h}\big) - \lambda_1||V\boldsymbol{h} - \boldsymbol{x}||_2 + \lambda_2 S\big(\boldsymbol{h}\big) \end{aligned} \tag{2}$$

with $c^*, \overline{c} \in C, c^*$ classifier outcome for the input $\boldsymbol{x}$ and $\overline{c} \neq c^*$.

Algorithm 2: Contrastive Explanation Maximization procedure

```
Input: data point $\boldsymbol{x} \in \mathbb{R}^d$, the number of antagonist classes $q$, the Oracle $\Omega$, a
       dictionary $V \in \mathbb{R}^{d \times k}$
Output: the encoding $\boldsymbol{h} \in \mathbb{R}^d$
$\boldsymbol{p} \leftarrow$ getClassProbabilities $(\boldsymbol{x}, \Omega)$;
$(c_1, c_2, \ldots, c_{q+1}) \leftarrow$ getBestClasses$(\boldsymbol{p}, q+1)$;
$\boldsymbol{h}_{expl} \leftarrow$ EMExplanationBuilder$(\boldsymbol{x}, c_1, \Omega, V)$;
for $i = 2$ to $q + 1$ do
    $\boldsymbol{h}_{anta}^{(i)} \leftarrow$ EMExplanationBuilder$(\boldsymbol{x}, c_i, \Omega, V)$;
end
return $\boldsymbol{h}_{expl}, \boldsymbol{h}_{anta}^{(2)}, \ldots, \boldsymbol{h}_{anta}^{(q+1)}$
```

4 Experimental Assessment

To test our framework, we chose as Oracle a convolutional neural network architecture, LeNet-5 [16], generally used for digit recognition as MNIST. We have trained the network from scratch using two different datasets: MNIST [16], and a subset of the e-MNIST dataset [7] composed of the first 10 lowercase letters. The model is learned using the Adam algorithm [15].

NMF with sparseness constraints [14] is used to determine the dictionaries. We set the number of atoms to 200, relying on PCA analysis which showed that the first 100 principal components explain more than 95% of the data variance. We construct different dictionaries with different sparsity values in the range $\gamma_1, \gamma_2 \in [0.6, 0.8]$ [14], then we choose the dictionaries having the best trade-off between sparsity level and reconstruction error. The dictionaries are determined by looking for a good trade-off between reconstruction error and sparsity level.

The atoms forming our explanations are selected by taking those with larger encoding values (i.e., those that are more "important" in the representation).

In Fig. 1 we show the proposed explanation from different inputs. The explanations are expressed in terms of two different set of atoms which in Sect. 3.3 we computed using h_{c*} and $h_{\overline{c}}$: the first one is the set of atoms which mostly contribute (in terms of weights) to the outcome of the Oracle, the second one the set of atoms which mostly contribute to a given constrastive outcome. For clarity, we chose the first five.

We can see that the atoms selected by h_{c*} provide elements which can be considered discriminative for the selected outcome, for example in Fig. 1a (red) EM selects many components which represent a diagonal line, showing that it is probably one of the main feature selected by the classifier to make its choice. In the second column (blue) we chose a contrast class (a "3") and we ask to the algorithm to make an explanation. We can see that the selected components which are mostly different and varied, showing that the given image, to be classified as a "3", should have also other characteristics, as the central horizontal line.

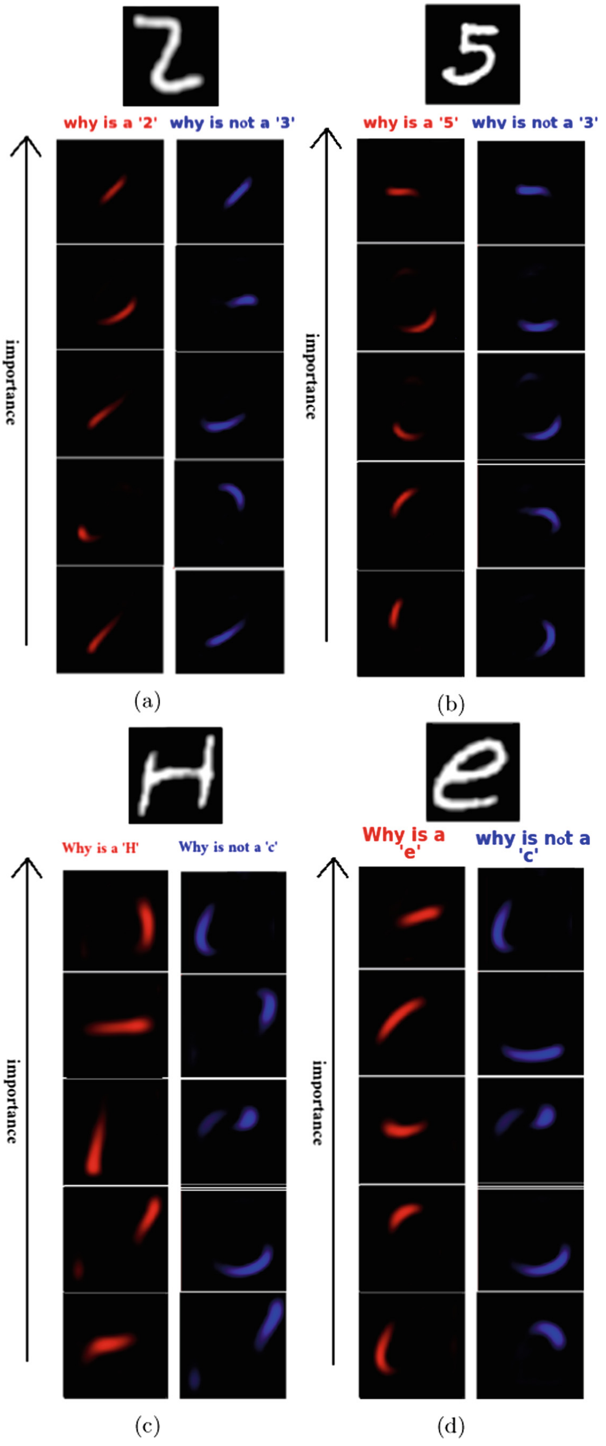

Fig. 1. Examples of direct and contrastive explanations. See discussion in Sect. 4 for more details (Color figure online)

Similar considerations can be made for the example shown in Fig. 1b, where the choice of a "five" can be motivated by the presence of the showed components (red), while in the blue column, we can notice the total absence of component on the left side, suggesting that the absence of a "left side" on the input image can be an explanation of why the given input has not been classified as a "3". in other terms, the input, to be classified as a "3", should have the visual components on the right side not relevant in terms of weights. In Fig. 1c the given input is correctly classified by the presence of the red component with high weights. The presence of the central line can be considered as the main discriminative feature between the outcome "H" and "c" (which is absent in the blue column). Similar considerations can be made for the input in Fig. 1d.

5 Conclusions

We proposed a model-agnostic framework to explain the answers given by classification systems. To achieve this objective, we started by defining a general explanation framework based on three entities: an Oracle (providing the answers to explain), an Interrogator (posing explanation requests) and a Mediator (helping Interrogator to interpret the Oracle's decisions). We propose a Mediator using known and established techniques of sparse dictionary learning, together with Interpretability ML techniques, to give a humanly interpretable explanation of a classification system outcomes. The proposed mediator can give explanation both in traditional and contrastive terms, since "why not?" questions are particularly relevant, from an ethical and legal viewpoint, to address user complaints about purported misclassifications and corresponding user requests to be classified otherwise. We tried our proposed approach by using an NMF-based scheme as sparse dictionary learning technique. However, we expect that any other technique that meets the requirements outlined in Sect. 3 may be successfully used to instantiate the proposed framework. The results of the experiments that we carried out are encouraging, insofar as the explanations provided seem to be qualitatively significant. Nevertheless, more experiments are necessary to probe the general interest of our approach to explanation. We plan to perform both a quantitative assessment, to evaluate explanations by techniques such as those proposed in [24], and a subjective quality assessment to test how do humans perceive and interpret explanations of this kind.

The proposed approach does not take so far into account factors such as the internal structure of the dictionary used. Accordingly, the present work can be extended by considering, for example, whether there are atoms that are sufficiently "similar" to each other or whether the presence in the dictionary of atoms which can be expressed as combinations of other atoms may affect the explanations that are arrived at.

References

1. Adadi, A., Berrada, M.: Peeking inside the black-box: a survey on explainable artificial intelligence (xai). IEEE Access **6**, 52138–52160 (2018)
2. Apicella, A., Isgrò, F., Prevete, R., Sorrentino, A., Tamburrini, G.: Explaining classification systems using sparse dictionaries. In: Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Special Session on Societal Issues in Machine Learning: When Learning from Data is Not Enough (2019)
3. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS One **10**(7), e0130140 (2015)
4. Bao, C., Ji, H., Quan, Y., Shen, Z.: Dictionary learning for sparse coding: algorithms and convergence analysis. IEEE Trans. Pattern Anal. Mach. Intell. **38**(7), 1356–1369 (2016)
5. Binder, A., Montavon, G., Lapuschkin, S., Müller, K.-R., Samek, W.: Layer-wise relevance propagation for neural networks with local renormalization layers. In: Villa, A.E.P., Masulli, P., Pons Rivero, A.J. (eds.) ICANN 2016. LNCS, vol. 9887, pp. 63–71. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-44781-0_8
6. Caruana, R., Lou, Y., et al.: Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1721–1730. ACM (2015)
7. Cohen, G., Afshar, S., Tapson, J., van Schaik, A.: EMNIST: an extension of MNIST to handwritten letters. arXiv e-prints arXiv:1702.05373 February 2017
8. Cooper, G.F., Aliferis, C.F., et al.: An evaluation of machine-learning methods for predicting pneumonia mortality. Artif. Intell. Med. **9**(2), 107–138 (1997)
9. Doran, D., Schulz, S., Besold, T.R.: What does explainable ai really mean? A new conceptualization of perspectives. arXiv preprint arXiv:1710.00794 (2017)
10. Dosovitskiy, A., Brox, T.: Inverting visual representations with convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 4829–4837 (2016)
11. Erhan, D., Bengio, Y., Courville, A., Vincent, P.: Visualizing higher-layer features of a deep network. University of Montreal 1341(3), p. 1 (2009)
12. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Comput. Surv. (CSUR) **51**(5), 93 (2018)
13. Hilton, D.J.: Conversational processes and causal explanation. Psychol. Bull. **107**(1), 65 (1990)
14. Hoyer, P.O.: Non-negative matrix factorization with sparseness constraints. J. Mach. Learn. Res. **5**(Nov), 1457–1469 (2004)
15. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. In: International Conference on Learning Representations, December 2014
16. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
17. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Advances in neural information processing systems, pp. 556–562 (2001)
18. Letham, B., Rudin, C., McCormick, T.H., Madigan, D., et al.: Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. Ann. Appl. Stat. **9**(3), 1350–1371 (2015)
19. Lipton, P.: Contrastive explanation. Roy. Inst. Philos. Suppl. **27**, 247–266 (1990)

20. Lipton, Z.C.: The mythos of model interpretability. Queue **16**(3), 30:31–30:57 (2018)
21. Mahendran, A., Vedaldi, A.: Understanding deep image representations by inverting them. In: CVPR, pp. 5188–5196 (2015)
22. Mensch, A., Mairal, J., Thirion, B., Varoquaux, G.: Dictionary learning for massive matrix factorization. In: International Conference on Machine Learning, pp. 1737–1746 (2016)
23. Miller, T.: Explanation in artificial intelligence: insights from the social sciences. Artif. Intell. **267**, 1–38 (2018)
24. Montavon, G., Samek, W., Müller, K.: Methods for interpreting and understanding deep neural networks. Digital Signal Process. **73**, 1–15 (2018)
25. Montavon, G., Lapuschkin, S., Binder, A., Samek, W., Müller, K.R.: Explaining nonlinear classification decisions with deep taylor decomposition. Pattern Recogn. **65**, 211–222 (2017)
26. Nguyen, A., Clune, J., Bengio, Y., Dosovitskiy, A., Yosinski, J.: Plug & play generative networks: Conditional iterative generation of images in latent space. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4467–4477 (2017)
27. Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T., Clune, J.: Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In: Advances in Neural Information Processing Systems, pp. 3387–3395 (2016)
28. Núñez, H., Angulo, C., Català, A.: Rule extraction from support vector machines. In: Esann, pp. 107–112 (2002)
29. Prevete, R., Apicella, A., Isgrò, F., Tamburrini, G.: Explaining the behavior of learning classification systems: a black-box approach. In: Proceedings of the 15th Conference of the Italian Association for Cognitive Sciences (2018)
30. Qin, Z., Yu, F., Liu, C., Chen, X.: How convolutional neural network see the world-a survey of convolutional neural network visualization methods. arXiv preprint arXiv:1804.11191 (2018)
31. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you? Explaining the predictions of any classifier. In: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144. ACM (2016)
32. Robnik-Šikonja, M., Kononenko, I.: Explaining classifications for individual instances. IEEE Trans. Knowl. Data Eng. **20**(5), 589–600 (2008)
33. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70, pp. 3145–3153. JMLR. org (2017)
34. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
35. Sturm, I., Lapuschkin, S., Samek, W., Müller, K.: Interpretable deep neural networks for single-trial eeg classification. J. Neurosci. Methods **274**, 141–145 (2016)
36. Szegedy, C., Zaremba, W., et al.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
37. 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
38. Zeiler, M.D., Taylor, G.W., Fergus, R.: Adaptive deconvolutional networks for mid and high level feature learning. In: ICCV, pp. 2018–2025 (2011)

39. Zhang, J., Bargal, S.A., et al.: Top-down neural attention by excitation backprop. Int. J. Comput. Vision **126**, 1084–1102 (2017)
40. Zhang, Q., Zhu, S.: Visual interpretability for deep learning: a survey. Front. Inf. Technol. Electron. Eng. **19**(1), 27–39 (2018)
41. Zintgraf, L.M., Cohen, T.S., Adel, T., Welling, M.: Visualizing deep neural network decisions: Prediction difference analysis. arXiv preprint arXiv:1702.04595 (2017)