Sparse dictionaries for the explanation of classification systems*

A. Apicella, F. Isgrò, R. Prevete, G. Tamburrini, and A. Vietri

Dipartimento di Ingegneria Elettrica e delle Teconologie dell'Informazione
Università degli Studi di Napoli Federico II, Italy
andrea.apicella@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, commercialization and use of machine learning systems pipeline is an important and challenging research problem. Crucial motivations to address this research problem can be advanced on both ethical and legal grounds. Notably, explanations of the decisions of machine learning systems appear to be needed to protect the dignity, autonomy and legitimate interests of people who are subject to automatic decision-making. 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. In this paper we discuss how the representational power of sparse dictionaries can be used to identify local image properties as main ingredients for producing humanly understandable explanations for the decisions of a classifier developed on the basis of machine learning methods.

Keywords: XAI · Explainable Artificial Intelligence · machine learning · sparse coding

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 behavior of which is often hard to interpret [15]. A crucial ML interpretability issue concerns the generation of explanations for an ML system behavior 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). Providing

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

XAI solutions to the ML explainability problem is important for many AI and computer science research areas: to improve intelligent systems design, 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 XAI solutions to the ML explainability problem is quite important from ethical and legal viewpoints as well. Learning systems are being increasingly used to make or to support decisions that are most significant for the life of persons, including career, court, medical diagnosis, insurance risk profiles and loan decisions. Thus, obtaining explanations for classifications and automatic decisions is arguably very important on ethical grounds, in order to respect and protect the dignity, autonomy and legitimate interests of people who are subject to automatic decision-making. On more properly legal grounds, it is sufficient to recall here that art. 22 of the European Union GDPR establishes the right of a person to contest an automatic decision and to address her complaint to the data protection officer who is in charge of the decision-making system. The data protection officer would be in a better position to evaluate these personal complaints, if he/she had an understanding of the reasons, if any, underlying the contested automatic decision. Moreover, in the case of repeated and undesired system behaviors, having good explanations of learning systems decisions can be very helpful to identify the sources of ethically unacceptable biases of learning systems, and to take those corrective actions that are impelled by codes of professional ethics.

Although some ML techniques come with reasonably interpretable mechanisms and Input/Output (I/O) relationships (e.g., decision trees), this is not the case for a wide variety of ML systems, whose processing and I/O relationships are often difficult to understand [14]. A ML system may have multiple sources of opacity for human bounded rationality, notably including the large numbers of features and ML parameters. As a consequence, the output of ML systems may depend from inner data representations and processing which escape full human understanding and interpretation. Indeed, in the ML system representation space, small differences or key features that cannot be easily made sense of in the framework of human classification strategies may play a decisive role for classification outcomes. Various senses of interpretability and explainability for learning systems have been distinguished and analyzed [7], and various approaches to overcoming their opaqueness are now being pursued [9, 22]. For example, in [19] a series of techniques for the interpretation of DNN are discussed, and in [16] 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 [27, 28], we focus on the issue of what it is to explain the behavior 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* [25].

Various model agnostic approaches have been developed to give *global* explanations exhibiting a class prototype which the input data can be associated to [9, 22, 27, 19]. These explanations are given in response to explanation requests that are usually expressed as why-questions: "Why were input data 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 impoverished 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 [25]

In this paper, we propose a model agnostic framework that returns local explanations of classifications based on *dictionaries* of local and humanly interpretable elements of the input. This framework can be functionally described in terms of a three-entity 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. The three-entity model is resumed in Figure 1. In this framework, local explanations are provided by a system (the Mediator) which does not coincide with the system which classifies inputs. A similar situation may occur in the human brain where, for instance, the visual system provides classifications and recognition of objects present in a visual scene, but the reasons why a given input is a recognised as a "cat" rather than, say, a ''dog", may involve other areas of the brain, including those storing and processing semantic memories. In this framework, the Mediator plays the crucial explanatory role, by advancing hypotheses on what humanly interpretable elements are likely to have influenced the Oracle output. 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 [18]. The dictionaries are obtained by a Non-negative Matrix Factorization (NMF) method [3, 14, 11], and the explanation is determined using an Activation-Maximization (AM) [9, 27] based technique, that we call *Explanation Maximization*.

The article is organized as follows: Section 2 briefly reviews related approaches, in Section 3 we present the overall architecture; experiments and re-

sults are discussed in Section 4, while Section 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 [23]) or rule-based systems [6, 5].

In the neural network context, recent surveys on explainable AI are proposed in [33, 24, 10, 1]. A significant attempt to explain in terms of images what a computational neural unit computes is found in [9] using the *Activation Maximization* method. AM-like approaches applied to CNN were proposed in [27, 17]. Additional attempts to give interpretability to CNNs were proposed in [31] and [8], where Deconvolutional Network (already presented by [32] as a way to do unsupervised learning) and *up-convolutional network* are proposed, while [22, 21] 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, [29] 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 [2, 4, 20] where Pixel-Wise Decomposition, Layer-Wise Relevance propagation ad Deep Taylor Decomposition are presented. [34] presents a work based on *prediction difference analysis* [26] where a features relevance vector is built which estimates how much each feature is "important" for the classifier to return the predicted class. In [25] , the model-agnostic explainer LIME is proposed, which takes into account the model behavior 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.

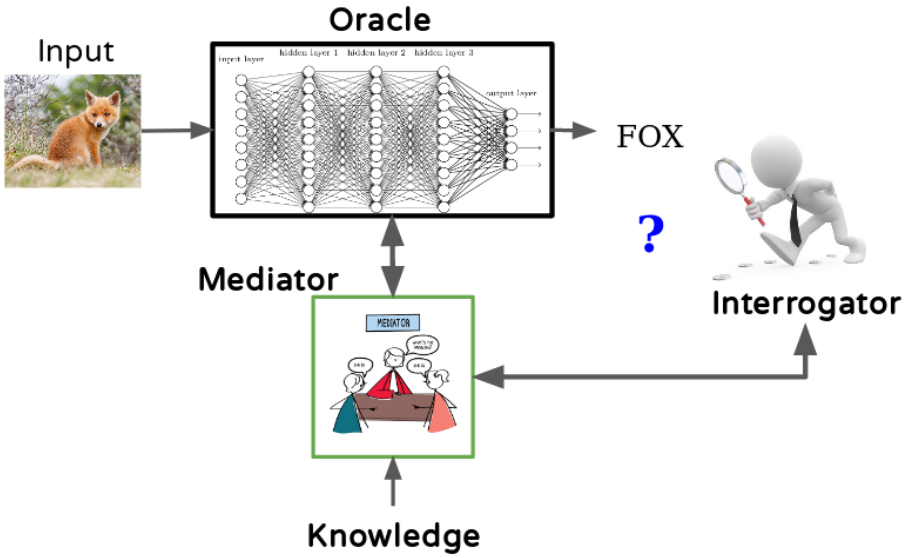

Fig. 1: The 3-entity proposed framework. See text for details.

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.

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 $D = \{(\boldsymbol{x}^{(1)}, c^{(1)}), (\boldsymbol{x}^{(2)}, c^{(2)}) \ldots, (\boldsymbol{x}^{(n)}, c^{(n)})\}$ 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 $X = (\boldsymbol{x}^{(1)}|\boldsymbol{x}^{(2)}|\ldots|\boldsymbol{x}^{(n)})$ 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 [14] with the additional sparseness constraint proposed by [11]; 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 Maximization

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 maximizes the answer of the model, it searches for the dictionary-based encoding $\boldsymbol{h}$ that maximizes 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 regularization term leads the algorithm to choose dictionary atoms that, with an appropriate encoding, form a good representation of the input, while the second regularization 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, $\forall i.1 \leq i \leq d$, 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 [11] that ensures both sparsity and non-negativity. The complete procedure is reported in Algorithm 1.

4 Experimental Assessment

To test our framework, we chose as Oracle a convolutional neural network architecture, LeNet-5 [13], generally used for digit recognition as MNIST. We have trained the network from scratch using two different datasets: MNIST [13], obtaining an accuracy of 98.86% on the test set, and Fashion-MNIST [30], obtaining

Algorithm 1: Explanation Maximization procedure

Input: data point $\boldsymbol{x} \in \mathbb{R}^d$, the output class $\hat{c}$,learned model Γ, a dictionary $V \in \mathbb{R}^{d \times k}$, λ_1, λ_2

Output: the encoding $\boldsymbol{h} \in \mathbb{R}^d$

$\boldsymbol{h} \sim U^d(0,1)$;
while $\neg$ *converge* **do**
  $\boldsymbol{r} \leftarrow V\boldsymbol{h}$;
  $\boldsymbol{h} \leftarrow \arg\max_{\boldsymbol{h}} \Pr\left(\hat{c}|\boldsymbol{r};\Gamma\right) - \lambda_1||\boldsymbol{r}-\boldsymbol{x}||_2$;
  $\boldsymbol{h} \leftarrow \text{proj}(\boldsymbol{h}, \lambda_2)$; ▷ proj$(\cdot,\cdot)$ is given by [11]
end
return $\boldsymbol{h}$;

an accuracy of 91.43% on the test set. The training set is composed of 50000 images, while the test set is composed of 10000 images; the model is learned using the Adam algorithm [12].

NMF with sparseness constraints [11] 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 $\gamma_1, \gamma_2 \in [0.6, 0.8]$ [11], then we choose the dictionaries having the best trade-off between sparsity level and reconstruction error. The dictionaries are determined by looking for a a good trade-off between reconstruction error and sparsity level.

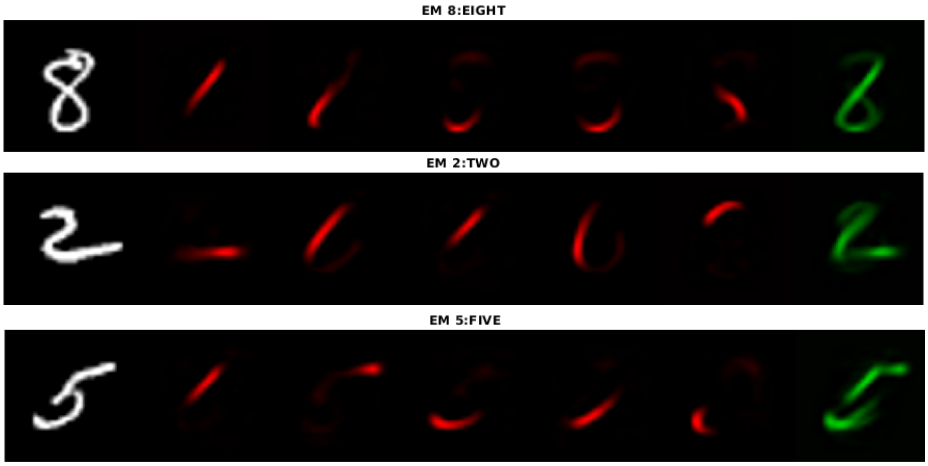

Fig. 2: Visual explanations obtained on three samples from the MNIST data set correctly classified by the Oracle. In red are the meaningful parts determined by the systems producing explanations. In green are the encodings of the input image obtained from the sparse dictionary.

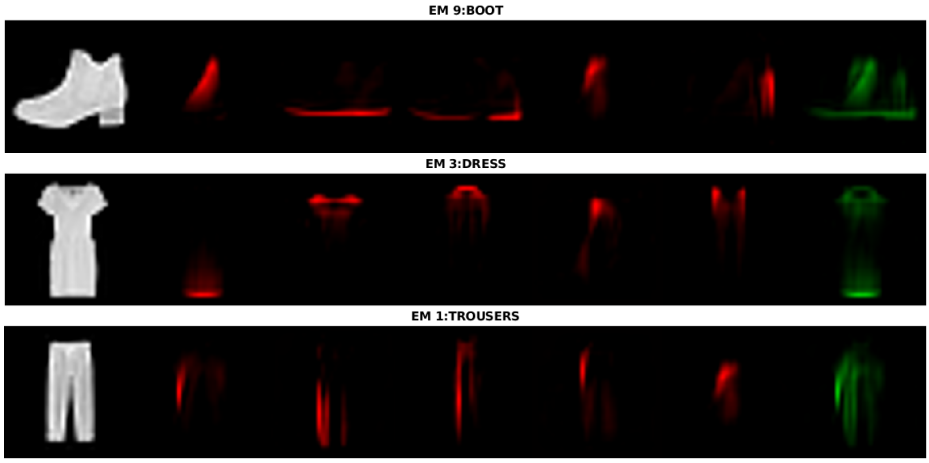

Fig. 3: Visual explanations obtained on three samples from Fashion-MNIST data set correctly classified by the Oracle. In red are the meaningful parts determined by the systems producing explanations. In green are the encodings of the input image obtained from the sparse dictionary.

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 figure 2 we show the atoms forming the explanation on three inputs for MNIST dataset on which the Oracle gave the correct answer. The chosen atoms seem to describe well the visual impact of the input numbers, by providing elements that appear to be discriminative, such as the crossed line and the bottom part for the "eight", a bottom straight line and a smoother upper part for the "two", and the straight upper part together with the curved bottom part for the "five". To probe empirically the impact of sparsity on this representation, we performed the same experiment using a dictionary with a very low sparsity (0.1), obtaining encondings without any preponderant value, thereby making it difficult to select appropriate atoms for explanation. In figure 3 we show the more "important" atoms obtained on three input images for the Fashion MNIST dataset, a *boot*, a *dress* and *trousers*, all of them correctly classified by the Oracle. Selecting the atoms with higher encoding values seems to give rise to representative parts of the selected input, returning parts that can be easily interpreted by an human interrogator, (e.g., the neck and the sole for the boot, the sleeves for the dress and the separation between the legs for the trousers).

As for MNIST, we performed the same experiment using a dictionary with low sparsity, ending up with results that are difficult to interpret.

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 ex-

planation 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. 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 Section 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 [19], 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. Another interesting direction of research concerns contrastive explanations, which enable one to answer "why not?" negative questions, by explaining why some given object was not given another classification, differing from the classification that the Oracle actually provided. One should be careful to note that "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.

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. 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)
3. Bao, C., Ji, H., Quan, Y., Shen, Z.: Dictionary learning for sparse coding: Algorithms and convergence analysis. IEEE transactions on pattern analysis and machine intelligence **38**(7), 1356–1369 (2016)
4. Binder, A., Montavon, G., Lapuschkin, S., Müller, K.R., Samek, W.: Layer-wise relevance propagation for neural networks with local renormalization layers. In: International Conference on Artificial Neural Networks. pp. 63–71. Springer (2016)
5. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., Elhadad, N.: Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1721–1730. ACM (2015)
6. Cooper, G.F., Aliferis, C.F., Ambrosino, R., Aronis, J., Buchanan, B.G., Caruana, R., Fine, M.J., Glymour, C., Gordon, G., Hanusa, B.H., et al.: An evaluation

of machine-learning methods for predicting pneumonia mortality. Artificial intelligence in medicine **9**(2), 107–138 (1997)
7. Doran, D., Schulz, S., Besold, T.R.: What does explainable ai really mean? a new conceptualization of perspectives. arXiv preprint arXiv:1710.00794 (2017)
8. Dosovitskiy, A., Brox, T.: Inverting visual representations with convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4829–4837 (2016)
9. Erhan, D., Bengio, Y., Courville, ., Vincent, P.: Visualizing higher-layer features of a deep network. University of Montreal **1341**(3), 1 (2009)
10. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM computing surveys (CSUR) **51**(5), 93 (2018)
11. Hoyer, P.O.: Non-negative matrix factorization with sparseness constraints. Journal of machine learning research **5**(Nov), 1457–1469 (2004)
12. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. In: International Conference on Learning Representations (12 2014)
13. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
14. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Advances in neural information processing systems. pp. 556–562 (2001)
15. Letham, B., Rudin, C., McCormick, T.H., Madigan, D., et al.: Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. The Annals of Applied Statistics **9**(3), 1350–1371 (2015)
16. Lipton, Z.C.: The mythos of model interpretability. Queue **16**(3), 30:31–30:57 (Jun 2018)
17. Mahendran, A., Vedaldi, A.: Understanding deep image representations by inverting them. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5188–5196 (2015)
18. Mensch, A., Mairal, J., Thirion, B., Varoquaux, G.: Dictionary learning for massive matrix factorization. In: Proceedings of The 33rd International Conference on Machine Learning. pp. 1737–1746 (2016)
19. Montavon, G., Samek, W., Müller, K.: Methods for interpreting and understanding deep neural networks. Digital Signal Processing **73**, 1–15 (2018)
20. Montavon, G., Lapuschkin, S., Binder, A., Samek, W., Müller, K.R.: Explaining nonlinear classification decisions with deep taylor decomposition. Pattern Recognition **65**, 211–222 (2017)
21. 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)
22. 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)
23. Núñez, H., Angulo, C., Català, A.: Rule extraction from support vector machines. In: Esann. pp. 107–112 (2002)
24. 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)
25. Ribeiro, M.T., Singh, S., Guestrin, C.: Why should i trust you?: Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD interna-

tional conference on knowledge discovery and data mining. pp. 1135–1144. ACM (2016)
26. Robnik-Šikonja, M., Kononenko, I.: Explaining classifications for individual instances. IEEE Transactions on Knowledge and Data Engineering **20**(5), 589–600 (2008)
27. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
28. Sturm, I., Lapuschkin, S., Samek, W., Müller, K.: Interpretable deep neural networks for single-trial eeg classification. Journal of neuroscience methods **274**, 141–145 (2016)
29. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
30. Xiao, H., Rasul, K., Vollgraf, R.: Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. CoRR **abs/1708.07747** (2017)
31. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European conference on computer vision. pp. 818–833. Springer (2014)
32. Zeiler, M.D., Taylor, G.W., Fergus, R.: Adaptive deconvolutional networks for mid and high level feature learning. In: Computer Vision (ICCV), 2011 IEEE International Conference on. pp. 2018–2025. IEEE (2011)
33. Zhang, Q., Zhu, S.: Visual interpretability for deep learning: a survey. Frontiers of Information Technology & Electronic Engineering **19**(1), 27–39 (2018)
34. Zintgraf, L.M., Cohen, T.S., Adel, T., Welling, M.: Visualizing deep neural network decisions: Prediction difference analysis. arXiv preprint arXiv:1702.04595 (2017)