Explaining Image Classifiers Generating Exemplars and Counter-Exemplars from Latent Representations

Riccardo Guidotti[1], **Anna Monreale**[2], **Stan Matwin**[3], **Dino Pedreschi**[2]
[1] ISTI-CNR, Pisa, Italy, riccardo.guidotti@isti.cnr.it
[2] University of Pisa, Italy, annam@di.unipi.it, pedre@di.unipi.it
[3] Dalhousie University and Polish Academy of Sciences, stan@cs.dal.ca

Abstract

We present an approach to explain the decisions of black box image classifiers through synthetic exemplar and counter-exemplar learnt in the latent feature space. Our explanation method exploits the latent representations learned through an adversarial autoencoder for generating a synthetic neighborhood of the image for which an explanation is required. A decision tree is trained on a set of images represented in the latent space, and its decision rules are used to generate exemplar images showing how the original image can be modified to stay within its class. Counterfactual rules are used to generate counter-exemplars showing how the original image can "morph" into another class. The explanation also comprehends a saliency map highlighting the areas that contribute to its classification, and areas that push it into another class. A wide and deep experimental evaluation proves that the proposed method outperforms existing explainers in terms of fidelity, relevance, coherence, and stability, besides providing the most useful and interpretable explanations.

Introduction

Today's automated decision systems for image classification are generally based on accurate machine learning techniques, such as deep neural networks. These models are recognized to be "black boxes" because of their opaque, hidden internal structure, whose complexity makes their comprehension for humans very difficult (Doshi-Velez and others 2017). Thus, there is an increasing interest in the scientific community in deriving explanations able to describe the behavior of a black box (Guidotti and others 2018), or explainable by design approaches (Kim and others 2016). In particular, explaining the reasons for a certain decision can be very important. For example, when dealing with medical images for diagnosing, how we can validate that an accurate image classifier built to recognize cancer actually focuses on the malign areas and not on the background for taking the decisions? Therefore, in this paper we investigate the problem of black box explanation for image classification.

In the literature, the problem is addressed by producing explanations through different approaches. On the one hand, gradient and perturbation-based attribution methods (Simonyan, Vedaldi, and Zisserman 2013; Shrikumar, Greenside, and others 2016) reveal saliency maps highlighting the parts of the image that most contribute to its classification. However, these methods are *model specific* and can be employed only to explain specific deep neural networks. On the other hand, *model agnostic* approaches can explain, also through a saliency map, the outcome of any black box (Ribeiro and others 2016; Guidotti and others 2019b). Agnostic methods may generate a local neighborhood of the instance to explain and mime the behavior of the black box using an interpretable classifier. However, these methods exhibit drawbacks that may negatively impact the reliability of the explanations. First, they do not take into account existing relationships between features (or pixels) during the neighborhood generation. Second, the neighborhood generation does not produce "meaningful" images since, e.g., some areas of the image to explain in (Ribeiro and others 2016) are obscured, while in (Guidotti and others 2019b) they are replaced with pixels of other images. Finally, transparent-by-design approaches produce prototypes from which it should be clear to the user why a certain decision is taken (Kim and others 2016). These approaches, however, cannot be used to explain an already trained black box: the transparent model has to be directly adopted as a classifier, possibly with limitations on the accuracy achieved.

We present ABELE, an Adversarial Black box Explainer generating Latent Exemplars (Guidotti et al. 2019). ABELE is a local, model-agnostic explanation method able to overcome the existing limitations of the local approaches by exploiting the latent feature space, learned through an adversarial autoencoder (Makhzani et al. 2015), for the neighborhood generation process. Given an image classified by a given black box model, ABELE provides an explanation for the reasons of the proposed classification. The explanation consists of two parts: *(i)* a set of *exemplars* and *counter-exemplars* images illustrating, respectively, instances classified with the same label and with a different label than the instance to explain, which may be visually analyzed to understand the reasons for the classification, and *(ii)* a *saliency map* highlighting the areas of the image to explain that contribute to its classification, and areas of the image that push

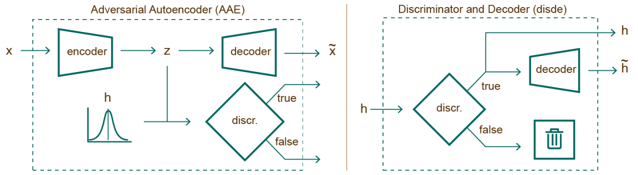

Figure 1: *Left*: Adversarial Autoencoder architecture.*Right*: Discriminator and Decoder ($disde$) module.

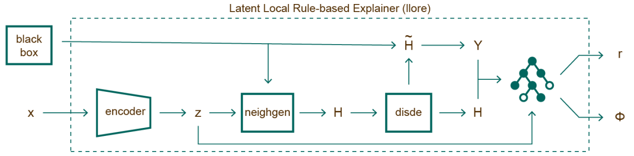

Figure 2: Latent Local Rules Extractor ($llore$) module.

it towards another label. Experiments empirically prove that ABELE overtakes state of the art methods by providing relevant, coherent, stable and faithful explanations.

Problem Formulation

We address the *black box outcome explanation problem* (Guidotti and others 2018). Given a black box model b and an instance x classified by b, i.e., $b(x) = y$, our aim is to provide an explanation e for the decision $b(x) = y$.

We focus on the black box outcome explanation problem for image classification, where the instance x is an image mapped by b to a class label y. In the following, we use the notation $b(X) = Y$ as a shorthand for $\{b(x) \mid x \in X\} = Y$. We denote by b a black box image classifier, whose internals are either unknown to the observer or they are known but uninterpretable by humans. Examples are neural networks and ensemble classifiers. We assume that a black box b is a function that can be queried at will.

Adversarial Black Box Explainer

ABELE (Adversarial Black box Explainer generating Latent Exemplars) is a local model agnostic explainer for image classifiers (Guidotti et al. 2019) Given an image x to explain and a black box b, the explanation provided by ABELE is composed of *(i)* a set of *exemplars* and *counter-exemplars*, *(ii)* a *saliency map*. Exemplars and counter-exemplars show instances classified with the same outcome as x, and with an outcome other than x, respectively. They can be visually analyzed to understand the reasons for the decision. The saliency map highlights the areas of x that contribute to its classification and areas that push it into another class. The explanation process involves the following steps. First, ABELE generates a neighborhood in the latent feature space exploiting an Adversarial Autoencoder (AAE) (Makhzani et al. 2015). Then, it learns a decision tree on that latent neighborhood providing local decision and counter-factual rules (Guidotti and others 2019a). Finally, ABELE selects and decodes exemplars and counter-exemplars satisfying these rules and extracts from them a saliency map.

Adversarial Autoencoders. AAEs are probabilistic autoencoders that aim at generating new random items that are highly similar to the training data. They are regularized by matching the aggregated posterior distribution of the latent representation of the input data to an arbitrary prior distribution. The AAE architecture (Fig. 1-left) includes an $encoder : \mathbb{R}^n \to \mathbb{R}^k$, a $decoder : \mathbb{R}^k \to \mathbb{R}^n$ and a $discriminator : \mathbb{R}^k \to [0, 1]$ where n is the number of pixels in an image and k is the number of latent features. Let x be an instance of the training data, we name z the corresponding latent data representation obtained by the $encoder$.

Encoding. The image $x \in \mathbb{R}^n$ to be explained is passed as input to the AAE where the $encoder$ returns the latent representation $z \in \mathbb{R}^k$ using k latent features with $k \ll n$.

Neighborhood Generation. ABELE generates a set H of N instances in the latent feature space, with characteristics close to those of z. Since the goal is to learn a predictor on H able to simulate the local behavior of b, the neighborhood includes instances with both decisions, i.e., $H = H_= \cup H_{\neq}$ where instances $h \in H_=$ are such that $b(\widetilde{h}) = b(x)$, and $h \in H_{\neq}$ are such that $b(\widetilde{h}) \neq b(x)$. We name $\widetilde{h} \in \mathbb{R}^n$ the decoded version of an instance $h \in \mathbb{R}^k$ in the latent feature space. The neighborhood generation of H ($neighgen$ module in Fig. 2) may be accomplished using different strategies ranging from pure random strategy using a given distribution to a genetic approach maximizing a fitness function (Guidotti and others 2019a). In our experiments we adopt the last strategy. After the generation process, for any instance $h \in H$, ABELE exploits the $disde$ module (Fig. 1-right) for both checking the validity of h by querying the $discriminator$ and decoding it into $\widetilde{h}$. Then, it queries the black box b with $\widetilde{h}$ to get the class y, i.e., $b(\widetilde{h}) = y$.

Local Classifier Rule Extraction. Given the local neighborhood H, ABELE builds a decision tree classifier c trained on the instances H labeled with the black box decision $b(\widetilde{H})$. Such a predictor is intended to locally mimic the behavior of b in the neighborhood H. The decision tree extracts the decision rule r and counter-factual rules Φ enabling the generation of *exemplars* and *counter-exemplars*. ABELE considers decision tree classifiers because: *(i)* decision rules can naturally be derived from a root-leaf path in a decision tree; and, *(ii)* counter-factual rules can be extracted by symbolic reasoning over a decision tree (Guidotti and others 2019a; Guidotti et al. 2019). Fig. 2 shows the process that, starting from the image to be explained, leads to the decision tree learning, and to the extraction of the decision and counter-factual rules. We name this module $llore$, as a variant of LORE (Guidotti and others 2019a).

Explanation Extraction. Often, e.g. in medical or managerial decision making, people explain their decisions by pointing to exemplars with the same (or different) decision outcome. We follow this approach and we model the explanation of an image x returned by ABELE as a triple $e = \langle \widetilde{H}_e, \widetilde{H}_c, s \rangle$ composed by *exemplars* $\widetilde{H}_e$, *counter-exemplars* $\widetilde{H}_c$ and a *saliency map* s. Exemplars and counter-

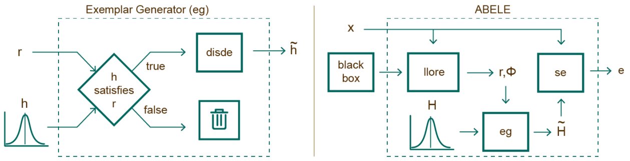

Figure 3: *Left*: (Counter-)Exemplar Generator (*eg*) module. *Right*: ABELE architecture.

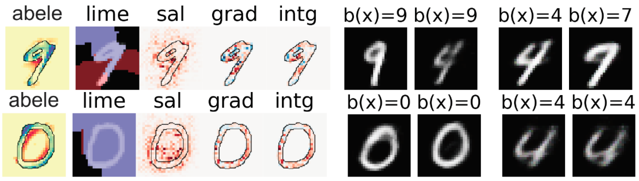

Figure 4: *Left*: Explanation by saliency maps comparison, *Right*: ABELE exemplars & counter-exemplars.

exemplars are images representing instances similar to x, leading to an outcome equal to or different from $b(x)$. Exemplars and counter-exemplars are generated by ABELE exploiting the *eg* module (Fig. 3-left). It first generates a set of latent instances H satisfying the decision rule r (or a set of counter-factual rules Φ), as shown in Fig. 2. Then, it validates and decodes them into exemplars $\widetilde{H}_e$ (or counter-exemplars $\widetilde{H}_c$) using the *disde* module. The saliency map s highlights areas of x that contribute to its outcome and areas that push it into another class. The map is obtained by the saliency extractor *se* module (Fig. 3-right) that first computes the pixel-to-pixel-difference between x and each exemplar in the set $\widetilde{H}_e$, and then, it assigns to each pixel of the saliency map s the median value of all differences calculated for that pixel. Thus, formally for each pixel i of the saliency map s we have: $s[i] = median_{\forall \widetilde{h}_e \in \widetilde{H}_e}(x[i] - \widetilde{h}_e[i])$.

Experiments

We ran experiments on three open source datasets[1]: `mnist`, `fashion` and `cifar10`, and we trained and explained Random Forest (RF), and Deep Neural Networks (DNN). We used 80% of the datasets for training the black boxes. 80% of the rest was used for training the AAE and the remaining 20% for evaluating the quality of the explanations. We compare ABELE against LIME and a set of saliency-based explainers collected in the `DeepExplain` package (DEX)[2]: Saliency (SAL) (Simonyan, Vedaldi, and Zisserman 2013), GradInput (GRAD) (Shrikumar, Greenside, and others 2016), IntGrad (INTG) (Sundararajan, Taly, and Yan 2017), etc. We compare the exemplars and counter-exemplars generated by ABELE against the prototypes and criticisms selected by MMD (Kim and others 2016) and by K-MEDOIDS.

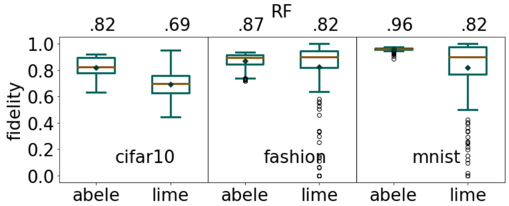

Figure 5: Box plots of *fidelity* (mean values on top).

Saliency Map, Exemplars and Counter-Exemplars. Before assessing quantitatively the effectiveness of the compared methods, we visually analyze their outcomes. We report explanations of the DNNs for the `mnist` dataset in Fig. 4 (left). The first column contains the saliency maps provided by ABELE: the yellow areas are common between x and the exemplars $\widetilde{H}_e$ (must remain unchanged to obtain the same label $b(x)$), the red areas are contained only in the exemplars and the blue ones are contained only in x (they can change without impacting the black box decision). With this type of saliency map we can understand that a *nine* may have a more compact circle and a *zero* may be more inclined. Moreover, besides the background, there are some "essential" yellow areas within the main figure that can not be different from x: e.g. the leg of the *nine*. The rest of the columns contain the explanations of the competitors: red areas contribute positively to the black box outcome, blue areas contribute negatively. For LIME, nearly all the content of the image is part of the saliency map, and the areas have either completely positive or completely negative contributions. The DEX methods return scattered red and blue points not clustered into areas. It is not clear how a user could understand the decision process of the LIME and DEX explanations. In addition, exploiting ABELE's explanations, in Fig. 4 (right) we show two exemplars and two counter-exemplars. We can notice how the label *nine* is assigned to images very close to a *four* but until the upper part of the circle remains connected, it is still classified as a *nine*. On the other hand, looking at counter-exemplars, if the upper part of the circle has a hole or the lower part is not thick enough, then the black box labels them as a *four* and a *seven*, respectively.

Interpretable Classifier Fidelity. We compare ABELE and LIME in terms of *fidelity*, i.e., the ability of the local interpretable classifier c of mimicking the behavior of b in the local neighborhood H. Fig. 5 shows that ABELE outperforms LIME with respect to the RF, while for the DNN LIME is slightly more faithful. However, for both RF and DNN, ABELE has a fidelity variance markedly lower than LIME.

Nearest Exemplar Classifier. Inspired by (Kim and others 2016), we test the goodness of exemplars and counter-exemplars by adopting a 1-Nearest Neighbor classifier (1-NN) trained on them. We generated n exemplars and counter-exemplars with ABELE, and we selected n prototypes and criticisms using MMD (Kim and others 2016) and K-MEDOIDS. Then, we employed a 1-NN to classify unseen instances. The classification accuracy reported in Fig. 6 is comparable among the methods. We observe that when n is

[1] We report here a selection of the results. The complete description of the results can be found in (Guidotti et al. 2019).

[2] Github code link: https://github.com/riccotti/ABELE.

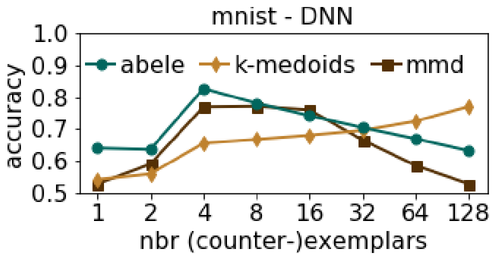

Figure 6: 1-NN (counter-)exemplar classifier accuracy.

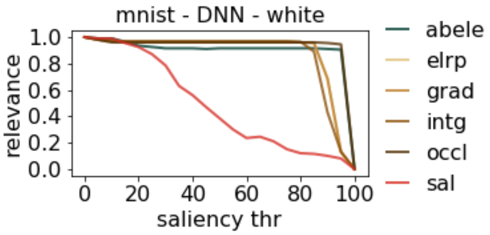

Figure 7: Relevance analysis (the higher the better).

low, ABELE outperforms MMD and K-MEDOIDS: just a few exemplars and counter-exemplars *generated* by ABELE are sufficient for recognizing the real label.

Relevance Evaluation. We evaluate the relevance of ABELE's saliency maps by partly masking the image to explain x. According to (Hara, Ikeno, and others 2018), although a part of x is masked, $b(x)$ should remain unchanged as long as relevant parts of x remain unmasked. We define the $relevance$ as the ratio of images in X for which the masking of relevant parts does not impact on the black box decision. The masking changes the pixels of x having a value in the saliency map smaller than the τ percentile of the set of values in the map. These pixels are substituted with the color *black*, *gray* or *white*. A low number of outcome changes means that the explainer successfully identifies *relevant* parts of the images. Fig. 7 shows the *relevance* for the DNN on `mnist` varying τ. ABELE is among the best performer masking with *white* or *gray*, while with *black*, ABELE's relevance is in line with those of the competitors.

Robustness Assessment. Since the stability of explanations is an important requirement for interpretability (Melis and others 2018), the analysis of the stability of interpretable classifiers (Guidotti and Ruggieri 2019) and explainers is crucial for gaining the trust of the user. We asses the *robustness* of the explanation methods through the local Lipschitz estimation (Melis and others 2018): a robust explainer must provide similar explanations to similar instances. Table 1 reports mean and standard deviation of the explainers' *robustness*. Our results show that LIME does not provide robust explanations, GRAD and INTG are the best performers, and ABELE performance is comparable to them. This high resilience of ABELE is due to the usage of AAE.

dataset	ABELE	GRAD	INTG	LIME
`cifar10`	$.57 \pm .10$	$.54 \pm .08$	$.53 \pm .11$	$1.9 \pm .25$
`fashion`	$.45 \pm .06$	$.49 \pm .10$	$.56 \pm .17$	$1.6 \pm .16$
`mnist`	$.38 \pm .03$	$.74 \pm .21$	$.78 \pm .22$	$1.5 \pm .14$

Table 1: Robustness for DNN (the lower the better).

Conclusion

ABELE is a local model-agnostic explainer using for the neighborhood generation the latent representations of an AAE. The explanation consists of exemplar and counter-exemplar images, labeled with the class identical to, and different from, the class of the image to explain, and by a a saliency map, highlighting the areas of the image contributing to its classification. Future research directions include extending ABELE to work on tabular data, text data and time series, and employing ABELE on a case study for explaining medical imaging tasks, e.g. radiography and fMRI.

Acknowledgments

This work is supported by the EC H2020 programme under the funding schemes: G.A. 654024 *SoBigData*, G.A. 78835 *Pro-Res*, G.A. 825619 *AI4EU* and G.A. 761758 *Humane AI*.

References

Doshi-Velez, F., et al. 2017. Towards a rigorous science of interpretable machine learning. *arXiv:1702.08608*.

Guidotti, R., et al. 2018. A survey of methods for explaining black box models. *ACM CSUR* 51(5):93:1–42.

Guidotti, R., et al. 2019a. Factual and counterfactual explanations for black-box decision making. *Intelligent Systems*.

Guidotti, R., et al. 2019b. Investigating neighborhood generation for explanations of image classifiers. In *PAKDD*.

Guidotti, R., and Ruggieri, S. 2019. On the stability of interpretable models. In *IJCNN*.

Guidotti, R.; Monreale, A.; Matwin, S.; and Pedreschi, D. 2019. Black box explanation by learning image exemplars in the latent feature space. In *ECML-PKDD*. Springer.

Hara, S.; Ikeno, K.; et al. 2018. Maximally invariant data perturbation as explanation. *arXiv:1806.07004*.

Kim, B., et al. 2016. Examples are not enough, learn to criticize! In *NIPS*.

Makhzani, A.; Shlens, J.; Jaitly, N.; Goodfellow, I.; and Frey, B. 2015. Adversarial autoencoders. *arXiv:1511.05644*.

Melis, D., et al. 2018. Towards robust interpretability with self-explaining neural networks. In *NIPS*.

Ribeiro, M., et al. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *KDD*, 1135–1144. ACM.

Shrikumar, A.; Greenside, P.; et al. 2016. Not just a black box: Learning important features through propagating activation differences. *arXiv:1605.01713*.

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2013. Deep inside convolutional networks. *arXiv:1312.6034*.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic attribution for deep neural networks. In *ICML*. JMLR.