Neurocomputing 562 (2023) 126884

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Role of locality, fidelity and symmetry regularization in learning explainable representations

Michele Ronco, Gustau Camps-Valls *

Image Processing Laboratory (IPL), Carrer del Catedrátic José Beltrán Martinez, 2, Paterna, València, 46980, Spain
Universitat de València, Av. de Blasco Ibáñez, 13, València, 46010, Spain

ARTICLE INFO

Communicated by T. Hu

Dataset link: https://github.com/IPL-UV/xAI-constrained-losses

Keywords:
Explainable AI
Deep learning
Interpretability
Attribution priors
Regularization

ABSTRACT

Despite their success deep neural networks still lack interpretability and are regarded as black boxes. This hampers a wider adoption in applications with societal, environmental or economical implications, and motivated a variety of techniques for explaining their outputs. Such explanations are however typically produced after model training so there is no guarantee that models learn faithful attributions, a goal they were not trained for. We evaluate the impact of different penalty terms in the loss function that promote explainable feature attributions, and that can be learned during training in an unsupervised way. We show that explainability-constrained models produce better saliency maps based on multiple metrics and tests. Regularizers imposing *locality*, *fidelity* and *symmetry* properties lead to the best performances in terms of MoRF and ROAR scores.

1. Introduction

In the last decade we have witnessed the rise of deep learning (DL) in a variety of different fields [1,2], spanning from computer vision to natural language processing. DL is gradually shaping and affecting our daily lives and promises to do so even more in the upcoming years. However, despite deep neural networks (DNNs) achieve impressive performance in multiple tasks, their decision process remains largely unknown. For this reason they are often referred to as black-box models. Even if, in some cases, we might not be interested in the underlying reasons behind DL's predictions (e.g. in some commercial applications), there are relevant domains where *understanding and not just fitting* becomes of pivotal importance: safety critical systems, health and natural sciences. Providing a trustworthy explanation together with an effective solution is necessary when applying DL to industrial environments where reliability is a must [3], in medicine where accountable diagnosis and management of patients' treatment is crucial [4–7], and physics where the main interest is in discovering general principles and laws that describe a process or even a class of processes [8–12]. Moreover, interpreting machine learning models can assist people in recognizing confounding factors in their training data or even dangerous and unfair relationships that models might learn unless we find ways of avoiding that. It is then necessary to make DL models more advantageous to us and safer by addressing the issue of their interpretability.

One possible solution consists in designing DL architectures which are more interpretable by construction. A recent work [13] introduced self-explainable neural networks which can be viewed as a non-linear generalization of linear models obeying a Lipshitz condition locally. Such a property can be satisfied thanks to the decomposition of the network architecture into a basis of functions or concepts and coefficients. Learning semantic concepts rather than just weights has been suggested also in [14], where convolutional neural networks (CNNs) are modified such that hidden-layer activations are whitened to align with predefined concepts. A related method introduced in [15] features a CNN for classifying images that uses prototypical aspects of each class rather than pixels directly. Despite these notable exceptions, the field of transparent DL architectures is still in its infancy and restricted to specific applications. More critically, it is not always clear how flexible these models are and whether they retain the performance of standard DL models. In general, fully transparent models lose by construction most of the complexity of artificial neural networks and, thus, tend to be closer to decision trees or linear models either in a local manner or in some parts of their structure.

The most popular approach is represented by explainable Artificial Intelligence (xAI) [16–24]. In this case DNNs are not modified but their predictions are accompanied by post-hoc explanations. Most methods for computing such explanations exploit the gradients of the trained network with respect to an instance and assign a score to all the input

* Corresponding author at: Universitat de València, Av. de Blasco Ibáñez, 13, València, 46010, Spain.
E-mail addresses: michele.ronco@uv.es (M. Ronco), gustau.camps@uv.es (G. Camps-Valls).

https://doi.org/10.1016/j.neucom.2023.126884
Received 9 September 2022; Received in revised form 17 February 2023; Accepted 3 October 2023
Available online 6 October 2023

features, eventually producing a map or saliency vector [25–30]. Since they are model agnostic techniques, there has been a widespread use of xAI in many applied domains. Remarkably, xAI helped us realizing that, in some cases, models do not actually learn meaningful patterns but rather rely on shortcuts and spurious correlations and end up producing biased predictions. For instance, a common mistake DL models commit is that of using background pixels to classify images [22]. Nonetheless, having multiple xAI techniques translates into multiple possible explanations to decide among, which can be troublesome. Moreover, it has been shown that some xAI do not pass simple sanity checks [31–34]. This raised some doubts on the reliability of saliency methods: either xAI methods fail to identify the correct attributions, or DNNs are right for the wrong reasons or, most likely, a combination of these two reasons. While there are extensive efforts to improve xAI techniques [16,17], much less attention has been devoted to finding ways of ensuring that DNNs learn explainable feature representations.

One could try to optimize some loss function that is believed to make the model more interpretable, for example, reducing complexity or improving the accuracy of feature attributions. The idea of optimizing saliency maps by constrained optimization was originally introduced in [35]: authors aimed to minimize not only the prediction error but also model's derivatives with respect to the input. This way, small changes in the input samples (e.g. due to noise) would not affect the error made by the model. The overall result was an improved model's weight distribution. However, their focus was not explicitly on the interpretability of DL models. Such an approach was rediscovered and popularized in [36,37] where gradients computed only with respect to irrelevant dimensions were regularized. These annotations were provided by domain experts knowing which features of x were (not) useful. Following such supervised schemes led to more robust models against adversarial attacks [37]. The incorporation of prior knowledge was also explored in [38] by minimizing the difference between contextual decompositions and ground-truthed explanations.

A fully unsupervised extension of Ross et al. [36] and Rieger et al. [38] was recently suggested in [39]: at each training step a perturbed input $\overline{x}$ is produced by masking the features with low gradient values, and then a penalty given by the Kullback–Leibler (KL) divergence between model's output at x and at $\overline{x}$ is minimized. Thus, unimportant features do not play a role in model's decision and explanations get sharpened. Another unsupervised approach was proposed in [40] where the penalty was the square difference between model's output and its local approximation given by LIME [22].

Attribution methods other than input gradients have been considered in [41,42]. The former [41] optimized the Wasserstein distance between the integrated gradients [28] of nearby samples, while Pillai and Pirsiavash [42] penalized the Grad-CAM [29] values over spurious features introduced as noise in input data. Different types of attribution priors (e.g. imposing smoothness) were considered in [43]. Finally, the increased computational cost due to attribution optimization was alleviated by proposing a subclass of neural networks in [44].

In this work, we greatly entend the types of possible explainability constraints in the loss function, which can be optimized, enabling us to enforce different kinds of desireble properties on the saliency maps. Moreover, unlike other works, we do not use ground-truth values (which are not available in most situations) for guiding the attributions scores during training but rather objective criteria suggested by axioms or formal results on *locality* and *equivariance* [45,46], and inspired by *fidelity* metrics for quantifying the goodness of saliency maps [34,47]. In this way, explanations are learned in an unsupervised fashion without the risk of enforcing confirmation biases. As main contributions of the work, we (1) introduce novel attribution priors that promote desirable properties of the saliency maps at training time; (2) showcase that these constraints for explanations, which are not satisfied by unconstrained models, can be learned via gradient descent without affecting model's accuracy; and (3) find that imposing locality, fidelity, and symmetry allows us to identify reliable features whose relevance is shared across different models even those whom have not been constrained for such specific priors. This suggests that, when adding constraints for the explanations, one also finds attributions which are more generalizable and less specific to a given model. Finally, we give empirical evidence of performance using several quantitative metrics, which had been previously introduced in the literature, and perform an extensive comparison of the commonalities and differences between the fature attributions obtained with (and without) explainability constraints. A simple qualitative inspection and visual comparison of prior-constrained saliency maps for a few test samples is also discussed. All the experiments can be reproduced by using the public repository at: https://github.com/IPL-UV/xAI-constrained-losses.

2. Axiomatic metrics for saliency map evaluation

Consider a typical classification problem, where we want to learn a function $f : \mathbb{R}^d \to \mathbb{R}^c$ on a subset of the input data $(X, Y) \sim D_{train} \subset D$, being c the number of classes, i.e. the target for a given instance x is $y \in \mathbb{R}^c$. The function or mapping f can be a neural network and $\hat{y} = f(x) = (f_1, \dots, f_c)$ the output of a `softmax` layer. Let us also assume that $x = (x_1, x_2, \dots, x_d) \in [0, 1]^d$. Then, we denote by $\phi(f, x) = (\phi_1, \phi_2, \dots, \phi_d) \in \mathbb{R}^d$ the *feature attribution* vector that, given a trained model f, associates a score to each of the d input dimensions of the sample x. A high score ϕ_k means that the kth feature is relevant for predicting y according to the mapping f and the algorithm used to compute $\phi(f, x)$, on the contrary low attributions identify less important or irrelevant input features. In this study, we will only consider *model-agnostic* $\phi(f, x)$ that do not depend on the specifics of the model (e.g. network's architecture), whose only requirement will be that of f being differentiable. The algorithm for attributions is often a function of the gradient of the predicted output with respect to the input, i.e. $\phi \propto \nabla_x f$ [25,26], which can be seen as a first-order approximation of the model near x. The function ϕ is often referred to as *saliency map* or more generally as *post-hoc explanation*. The explanation induces a natural ordering in the input space, and we call $S_i = S(\phi_i)$ the ordering of the features which are sorted by $S(\cdot)$ in descending order of explanation score.

A lack of ground-truth explanations and limited theoretical understanding of both DNNs and xAI have led to a growing class of metrics [31,34,47–52]. Each of them measures a different property that a correct explanation should have according to empirical criteria or formal axioms. A metric $M(x, f, \phi)$ quantifies how accurate is the explanation ϕ. We here consider the following four.

Most relevant first out. The concept of *fidelity* of the explanation was introduced in [47]. If $\phi(f, x)$ encodes reliably the model prediction at x then the progressive removal of the most important features should deteriorate the accuracy of f at x. The Most Relevant First out (MoRF) procedure consists in computing the ratio between $f(x)$ and $f(x_{i \in S_k = 0})$:

$$\mathrm{MoRF}_{S_k} = 1 - \left\langle \frac{f_m(x_{i \in S_k = 0})}{f_m(x)} \right\rangle, \tag{1}$$

where $\langle \cdot \rangle$ stands for the average over all test samples, $f_m = \max(f(x))$ is the component of the `softmax` according to the prediction on the original instance, and $x_{i \in S_k = 0}$ is the perturbed input where the first S_k most relevant features have been inhibited. The ordering S_k is specified by a given xAI saliency map $\phi(f, x)$. Features can be removed either by setting them to 0 or the average or any other 'neutral' value. By varying the amount k of removed dimensions in (1), one obtains a perturbation curve that describes how the prediction for the class c changes as more features are cancelled out, from the most to the least important. The Area Under the MoRF Curve ($\mathrm{AUC}_{\mathrm{MoRF}}$) provides a summarizing quantitative measurement of the correctness of ϕ: the higher the $\mathrm{AUC}_{\mathrm{MoRF}}$ the more accurate the estimate of input feature importance. Let us stress that such a correctness is according to the

model and does not refer to human judgement nor to causal notions of feature importance. In other words, it is a measure of how well the xAI method captures the characteristics that are actually important for the model itself to make a prediction.

Faithfulness. Another metric that relies on perturbing the input x according to the attributions ϕ is *faithfulness* [13,34], which equals the Pearson's correlation coefficient $\rho(\cdot,\cdot)$ between the attributions of S_k features and the difference between the output at x and at $x_{i\in S_k=0}$:

$$F = \left\langle \rho(\vec{\phi}_S, \vec{f}_l(x) - \vec{f}_l(x_{i\in S_k=0})) \right\rangle, \tag{2}$$

where $\vec{\phi}_S = (\phi_{S_1}, \phi_{S_2}, \dots, \phi_{S_d})$ if $k = 1$, i.e. if we perturb the features one by one, otherwise $\vec{\phi}_S \in \mathbb{R}^{h<d}$ for $1 < k \leq d$, while $f_l(x)$ is the element of the `softmax` corresponding to the ground-truth label l. The perturbed input can be obtained by setting S_k features to 0 or to some other appropriate value. A high F means that the attribution scores ϕ_k correctly reflect model's behaviour, since there is a high correlation between the changes in the attribution and those in the predictions of the model.

Complexity. It is generally believed that the effectiveness of neural networks relies on the fact that the relevant information is encoded in a manifold of much lower dimensionality than that of the input space $\mathbb{R}^d$ [2]. Therefore, 'good' feature attributions ϕ_k should be sparse and concentrated in a few input dimensions. Thus, *complexity* [51] or entropy can be used for comparing different $\phi(f,x)$:

$$C = -\sum_{i=1}^{d} \mathbb{P}_\phi(i) \ln(\mathbb{P}_\phi(i)), \quad \text{where} \quad \mathbb{P}_\phi(i) = \frac{|\phi_i|}{\sum_{i=1}^{d} |\phi_i|}. \tag{3}$$

The lower C the less complex the explanation and, possibly, the more faithful to the latent representation learned by f. According to Bhatt et al. [51], we should favour those attributions $\phi(f,x)$ which are concentrated in a few input features. However, note that the minimum value for C is obtained when $\mathbb{P}_\phi(i) = 1$ i.e. $\phi_i \neq 0$ for only one feature i. Since typically the effective dimension is greater than 1, an attribution such that $C = 0$ will not be a correct explanation, and thus in general the best value for C will depend on the specific problem [53].

Remove and retrain. Most of the metrics for saliency maps rely on some form of perturbation of the test data. However, as DNNs can easily deteriorate in presence of distribution shifts [2,54], it is hard to determine whether the model errors in both (1) and (2) are due to the fact that important features have been identified by $\phi(f,x)$ and removed or just a consequence of the fact that the perturbed input is an out-of-sample for which the model struggles to make the correct prediction [31,52]. To obtain a fair evaluation of feature attributions the RemOve And Retrain (ROAR) test was introduced [31]. First, one computes $S_i = S(\phi_i(f,x))$ for all test and train instances, using the preferred attribution method to get ϕ. A perturbed (or masked) dataset is constructed by taking $x_{i\in S_k=0}$ in both train and test data. Then, a new model is trained over the perturbed data and its performance is evaluated on the perturbed test set. The procedure is repeated for different levels of degradation by increasing the number of perturbed features k and thus a perturbation curve is obtained. Owing to L-curve theory, the steepest the curve the better the explanation $\phi(f,x)$ is considered. Notice that, even if ROAR corrects for distribution shift effects, in each iteration a different model is trained and thus it is difficult to establish to what extent these new models reflect the original one.

3. Explanation-constrained loss functions

Training the classifier $f = f(X,w)$ generally involves minimizing a loss function of the form:

$$L(w,X,Y) = -\sum_{n=1}^{N}\sum_{k=1}^{c} y_{nk} \log(\hat{y}_{nk}) + \lambda\Omega(w) = \mathcal{L}_{CE}(Y,w) + \lambda\Omega(w), \tag{4}$$

where $\mathcal{L}_{CE}$ is the cross-entropy term and $\Omega(w)$ is the regularization term, N is the total number of training samples and c the number of classes in the target variable y_{nk}. The regularization term usually takes the form of either the ℓ_1 or ℓ_2 norm of model weights w. Following [36,37], we here generalize the regularization term as:

$$\Omega = \Omega(\phi(f,X)), \tag{5}$$

where now Ω imposes a constraint on the feature attribution $\phi(f,X)$. In this way desirable (explanatory) properties can be enforced for differentiable Ω. The penalty Ω is called *attribution prior* or *explainability constraint*. Following [55], one can reformulate the learning problem in Eq. (4) as:

$$p(M|X) \propto \max_{M\in\mathcal{M}} p(X|M)p(M) \tag{6}$$

being M a given model, $p(M|X)$ the posterior probability distribution, $p(X|M)$ the likelihood and $p(M)$ the prior over all possible M in the set $\mathcal{M}$. The prior $p(M)$ can be used to impose properties on the class of models such as sparsity. For instance a Gaussian prior on w is equivalent to ℓ_2 regularization. Explainability constraints can be seen in the same way as priors that enforce some degree of interpretability (which will be defined below in different ways through different types of penalties) on the models. Finding the explicit form of $p(M)$ for many $\Omega(\phi(f,X))$ can be highly non-trivial and it is part of current research (see e.g. the discussion in [43]).

Firstly, we review two attribution priors already appeared in the literature, namely vanilla gradient regularization [36] and the smoothness prior [43]. Then, with the goal of improving the quality of explanations, we suggest four new constraints to be optimized during training: fidelity, locality, symmetry and consistency. These additional regularization terms, which have not considered before, impose different properties on the explanations and, as proven in the experiments, are able to improve the interpretability of the saliency maps obtained in previous studies in terms of MoRF and ROAR scores. This also shows the flexibility of the proposed approach and demonstrate the possibility of introducing a variety of explainability constraints even beyond those considered here. In this study we assume $\phi(f,X) = \nabla_X f$ [25–27], unless otherwise specified.

Regularization. Inspired by regularization theory, one can extend it to model's gradients as done in [36]:

$$\Omega(\phi(f,X)) = \|\nabla_X f\|_2 = \sum_{n=1}^{N}\sum_{i=1}^{d}\left(A_{ni}\frac{\partial}{\partial x_{ni}}\sum_{k=1}^{c}\log(\hat{y}_{nk})\right)^2, \tag{7}$$

which is known as *input gradients regularization* or *double backpropagation* as it requires computing the quantity $\nabla_w\nabla_X f$ during training. Here N is the number of training samples and d the number of features (e.g. the number of pixels in a black and white image). Notice that, if the model is linear, then (7) is equivalent to ℓ_2 (or ℓ_1) regularization. In the simplest case the matrix A_{ni} reduces to a constant scalar [35] and, thus, large values for all gradients are discouraged. As a result, small changes in the input x do not affect the model output thereby improving generalization [35]. Alternatively, when supervised annotations are available, one can penalize specifying non-informative components of A_{ni} [36,37].

Locality. We here propose a possible extension of input gradients regularization in the form of a *locality* prior, particularly useful in image or language processing. To avoid focusing on spurious or noisy dimensions, we suggest the following penalty term:

$$\Omega(\phi(f,X)) = -\sum_{n=1}^{N}\sum_{i=1}^{d}\left(A_{ni}\log(\phi_{ni}) + (1-A_{ni})\log(1-\phi_{ni})\right), \tag{8}$$

where $\phi \in [0,1]^{N\times d}$ are the normalized input gradients and $A_{ni} = 0$ for all the irrelevant dimensions. If x has a zero background then $A_{ni} \equiv x_{ni}$, otherwise A can be derived from a segmentation mask or clustering. In many cases, especially when x is an image, the signal is concentrated in

a few input dimensions and the remaining pixels are just background or noise. There are some popular examples in which trained models end up relying fallaciously on background features and eventually produce biased predictions [22]. The above introduced penalty forces the model to neglect such features.

Smoothness. In [43], inspired by total variation concepts in image processing, the authors imposed a *smoothness* constraint by penalizing the total difference between attributions of adjacent pixels:

$$\Omega(\phi(f,X)) = \sum_{n=1}^{N}\sum_{i,j=1}^{h,w} |\phi^n_{i+1,j} - \phi^n_{i,j}| + |\phi^n_{i,j} - \phi^n_{i,j+1}|, \tag{9}$$

that assumes a zero-mean Laplace prior $p(M)$ on the distribution of differences of nearby attributions (see again [43] and references therein). Here h and w stand for the height and width in the input image respectively. In this way pixels which are close to each other spatially will be assigned similar attribution scores by the model.

Fidelity. Removing informative features according to $\phi(f,X)$ should make the output of the classifier change, cf. Section 2. We introduce a suitable attribution prior for enforcing the fidelity of explanations:

$$\Omega(\phi(f,X)) = \max\left(0, \left\langle d(f(x), f(\overline{x}))\right\rangle - \epsilon\right), \quad d(f(x), f(\overline{x})) = \frac{f\cdot\overline{f}}{\|f\|\|\overline{f}\|},$$
$$\overline{x} = x\odot(1-\phi), \tag{10}$$

where $\langle\cdot\rangle$ is the average over training X or a batch, $d(\cdot,\cdot)$ is the angular distance or cosine similarity (where $f\cdot\overline{f}$ stands for the scalar product between the two vectors), $\odot$ denotes the Hadamard or component wise product, ϕ are the normalized input gradients, and hyperparameter $\epsilon \geq 0$ controls the fidelity between $f(x)$ and $f(\bar{x})$: e.g. for $\epsilon = 0$ the fidelity constraints forces $f(x)$ and $f(\overline{x})$ to be orthogonal, i.e. a complete degradation of the model output when the salient features get cancelled or attenuated. The attributions ϕ are normalized in such a way that the masked or degraded input $\overline{x}$ (which takes values in the range $[0,1]$) contains (close to) zero signal in those dimensions with higher importance $\phi \approx 1$. Different functions for the distance $d(\cdot,\cdot)$ could be easily implemented, e.g. the Euclidean distance, depending on the form of the output of the model $f(x)$, as well as different losses $\Omega(\phi(f,X))$ to enforce some degree of fidelity in the explanations (see (1)).

Consistency. A trustworthy and interpretable explanation should be general enough to encode patterns common to all (or most of the) elements in a given class rather than traits specific to each instance. Arguably, close by instances in the input space that belong to the same class should have similar explanations. We here translate such property of *consistency* [56,57] into the following attribution prior:

$$\Omega(\phi(f,X)) = \sum_{k=1}^{c}\sum_{n,m=1}^{N_k} \frac{1-d(\phi_n,\phi_m)}{1-d(x_n,x_m)}, \tag{11}$$

where N_k are all the instances in a batch (or in the whole training set) predicted to belong to class k, and d is a similarity measure. As in (10) we choose the cosine distance but the extension to other types of similarity measures is immediate. The above regularization terms is such that similar instances (i.e. with cosine similarity close to 1), which are predicted to be in the same target class, should also have similar attributions (i.e. $d(\phi_n,\phi_m)\approx 1$). In this way the model is required to be consistent in its explanations.

Symmetry. There is a growing evidence that DNNs exploit the symmetries in the input data distribution. This is extensively used in data augmentation techniques [58–60], but also as a guiding principle for designing new architectures that encode a class of symmetry transformations by construction [45,61,62]. Following this perspective, a correct explanation ϕ should identify a subset of semantic features that have good transformation properties [46], i.e.:

$$x'_k = T_{ki}x_i \xrightarrow{f} \phi'_k = T_{ki}\phi_i, \tag{12}$$

being T_{ki} a representation of a given transformation operator, e.g. the generator of translations or rotations in a plane. Whether such a property is satisfied depends on the model f as well as on the attribution method for computing ϕ. We impose (12) by introducing the *symmetry* attribution prior:

$$\Omega(\phi(f,X)) = 1 - \left\langle |d(\phi'(x,f), \phi(x',f))|\right\rangle, \tag{13}$$

with x' and ϕ' as defined in (12) and we consider rotation transformations. In short, one requires that the transformed attributions are equal to the attributions of the transformed input.

4. Experimental results

In all our experiments we implement shallow CNNs made of two convolutional blocks and one fully connected layer with `softmax` activations and restrict the training between 6 to 8 epochs, in this way we reduce the additional computational cost carried by the penalty terms and do not compromise the analysis of expressiveness by using potentially overfitted models. The regularization parameters λ are obtained through grid search in the range $[0.05, 1.5]$ for each constraint. All models are implemented in PyTorch and trained with Adam [63]. Further details on the architectures, attribution priors and hyperparameters, as well as additional experiments can be found in Appendices A.1–A.3.

Firstly, we notice that all trained models can reach near to state of art performance on MNIST [64], even if the training accuracy can have a slightly less steep increase depending on the explainability constraint (see Fig. 1a). Confirming the early findings in [35], all constraints improve the distribution of the learned weights to different degrees as compared to the unconstrained model (i.e. *baseline*). This observation holds in particular for vanilla *input gradients regularization* (7) and *smoothness* (9) (see the second plot in Fig. 1a). Fig. 1b shows that traditional training without regularization does not optimize any of the feature attribution properties. This suggests that the trained models have learned distinct patterns for making predictions.

We claim that, even if looking at the accuracy only, explainability-constrained models are almost indistinguishable from the *baseline* model, yet the former lead to more interpretable and reliable explanations. In order to assess whether attribution priors truly enable models to better discriminate between more and less important features, we use different constraints (cf. Section 3) and metrics (cf. Section 2). Results are shown in Fig. 2, and Table 1 summarizes the metrics for the saliency maps.

From the first plot in Fig. 2a we can distinguish two groups of models, i.e. models with *locality* (8), *fidelity* (10) or *symmetry* (13) constraints which have a high MoRF score (1) and the others with lower scores (see also Table 1). Moreover, if now we recalculate this metric for all the models but using e.g. the feature attributions $\phi = \nabla_x f_{\text{symmetry}}$ from the model with symmetry penalty to decide in which order S_k the input features should be removed, the MoRF curve gets steeper for all the models including the baseline (see the middle plot in 2a). The opposite behaviour is observed if we employ the ranking from the unconstrained model instead, i.e. $S = S(\phi(f_{\text{baseline}}, x))$ (see last plot in 2a). This means that the explanations obtained from the model with the symmetry regularization (13) are more trustworthy for all the other models too, and conversely traditional training results in less reliable attributions regardless of the considered model. Similar conclusions hold for locality and fidelity, see Appendix A.2.

Attribution priors allow us to improve also the faithfulness (2) score (especially input gradient regularization and smoothness) and the metric for complexity (3), see Table 1. In particular, it is interesting to note that the lowest complexity is achieved thanks to the symmetry penalty. This can be explained by the fact that requiring the correct transformation properties under rotations reduces the number of allowed coordinates thereby eliminating the spurious input features. Such features can be for instance background pixels or objects in images or also corrupted patches. In general they can be defined as all those

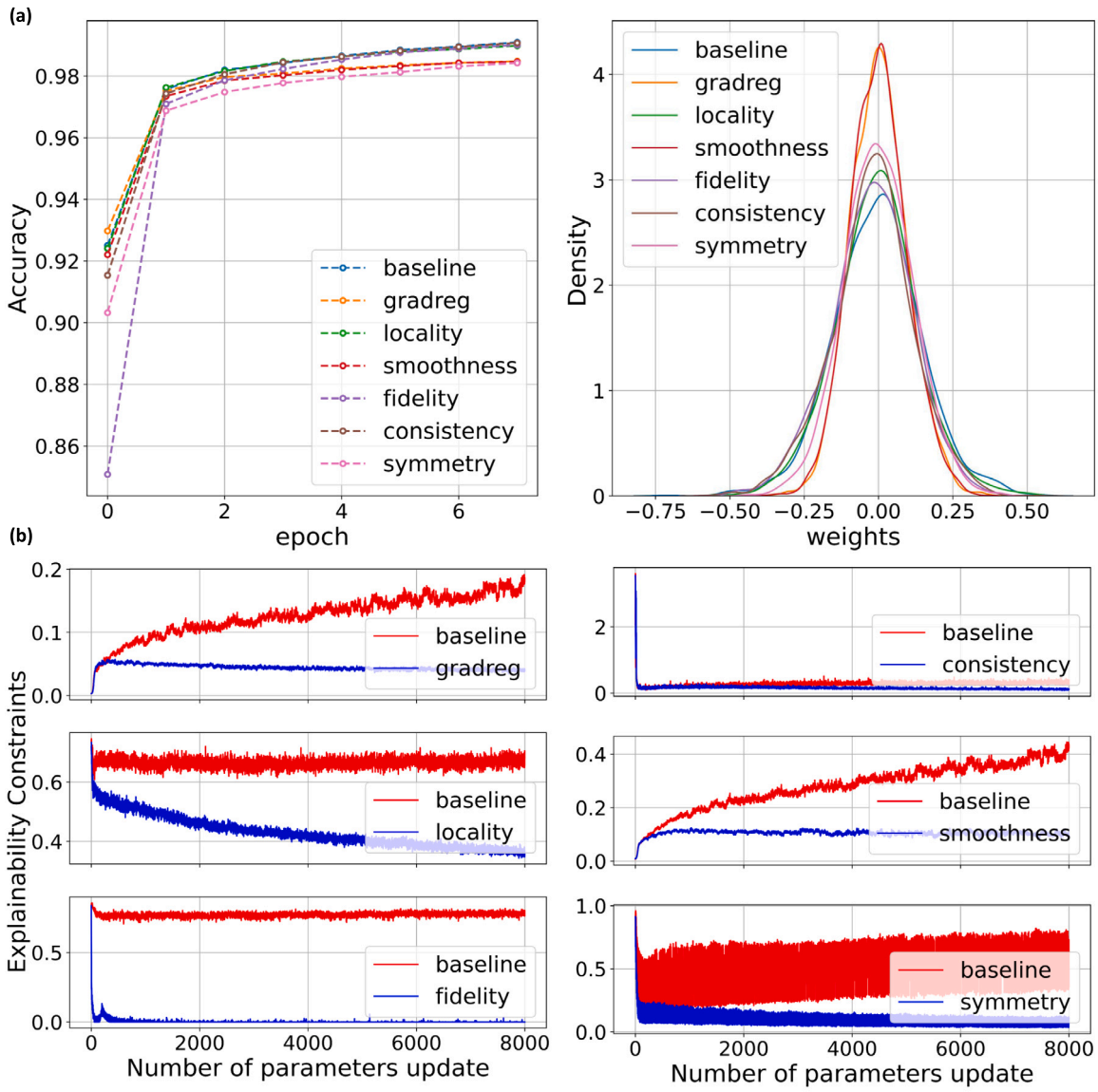

Fig. 1. (a) Training accuracy per epoch for all models. Explainability constraints generally slow down learning, but similar accuracy is reached eventually. All constraints act as weight regularizers, smoothness and gradient regularization leading to especially reduced weight variances. (b) Explainability constraints versus the training steps. None of the desired properties encoded in the attribution priors get automatically optimized by the baseline model trained without penalty terms. From left to right and top to bottom, in the order we display the following regularization terms: input gradients regularization (Eq. (7)), consistency constraint (Eq. (11), locality constraint (Eq. (8)), smoothness constraint (Eq. (9)), fidelity constraint (Eq. (10)) and symmetry constraint (Eq. (13)). The red curve represents the trend of the corresponding regularization term in the baseline model (i.e. without any regularization) during training, while the blue curve is the same quantity computed with the model that contains that specific constraint and, thus, by construction decreases over the epochs.

Table 1
Metrics for saliency maps.

Constraint	MoRF	Faithfulness	Complexity	ROAR	Gaussian	Block
Baseline	0.696	0.800	4.40	0.835	0.643	9.79
GradReg	0.658	**0.909**	4.33	0.767	0.660	9.89
Locality	**0.846**	0.707	4.35	0.753	0.388	9.53
Smoothness	0.688	0.894	4.33	0.755	**0.818**	9.76
Fidelity	0.827	0.797	4.30	0.724	0.470	**9.93**
Consistency	0.626	0.822	4.38	0.844	0.523	9.77
Symmetry	0.836	0.820	**4.23**	**0.666**	0.311	9.13

features which might have some degree of correlation with the target (and, thus, unconstrained models could pick them) in the training set (e.g. due to noise or biases), but do not have a predictive power on unseed test samples.

A common benchmark for the accuracy of explanations is the ROAR test [31]. We apply it to the saliency maps from all the explainability-constrained models which are compared against the baseline (see the first plot in 2b). To speed up the test, for each fraction of removed features we train a model without regularization over the perturbed train data for a total of 6 epochs. Thus, the curves in Fig. 2b are obtained with different attributions ϕ but the retrained model is of the same type for all of them. The most reliable explanation is the one that most degrades the test accuracy for a given fraction of removed features. We can see that, as expected, the degradation of accuracy over the test set decreases monotonically for all models. The steepest curve corresponds to the model (and, consequently, explainability constraint) whose feature attributions are more precise and reliable according to the ROAR score (see e.g. [31,65]). Baseline attributions are clearly improved and, in particular, attributions from symmetry and fidelity

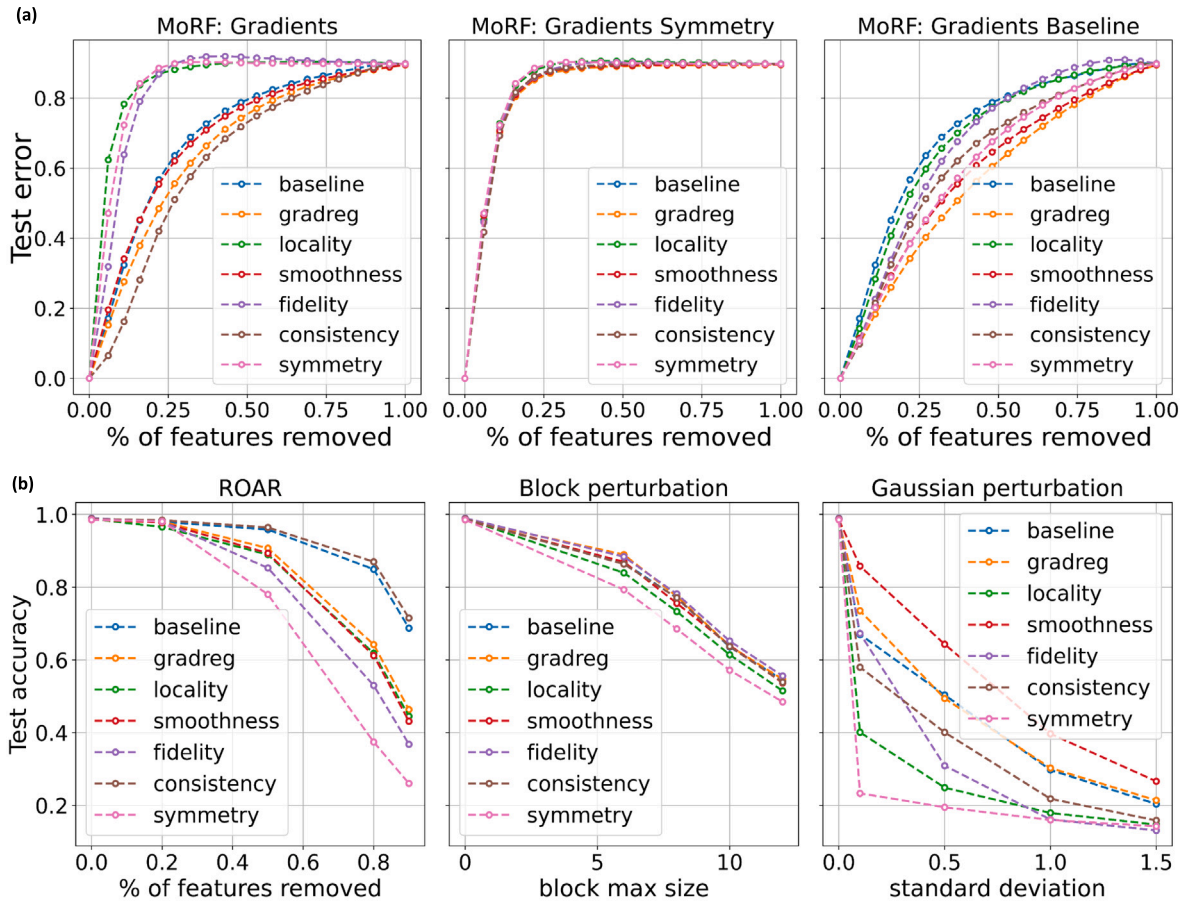

Fig. 2. Test error (a, top row) and accuracy (b, bottom row) as a function of the rate of removed features or the amount of perturbation injected for different constraints (baseline, GradReg, locality, smoothness, fidelity, consistency and symmetry) and metrics (MoRF, faithfulness, complexity, ROAR, Gaussian and block perturbations).

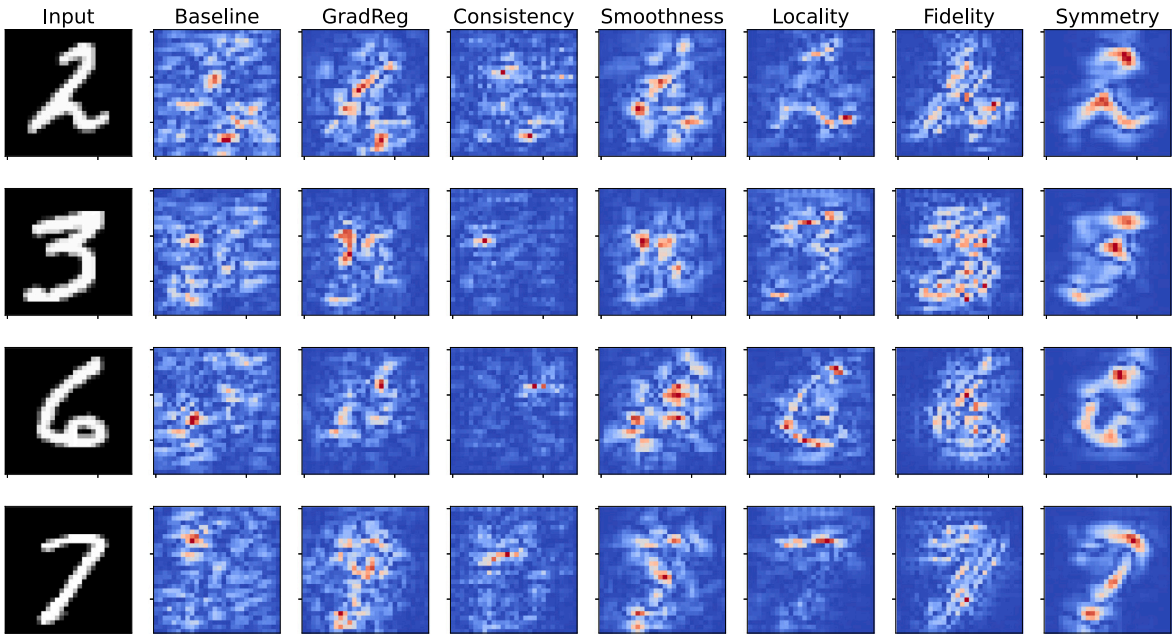

Fig. 3. Saliency maps for some input images. One can notice that explainability-constrained models produce more interpretable attributions with respect to the noisy saliency from the baseline. In the above images pixels coloured in blue correspond to zero or negative attributions while those in the red scale have positive attributions. We remind that the higher the attributions the more important are the corresponding features in predicting the output, while features with zero or negative attributions are either inactive or push the model towards a wrong class.

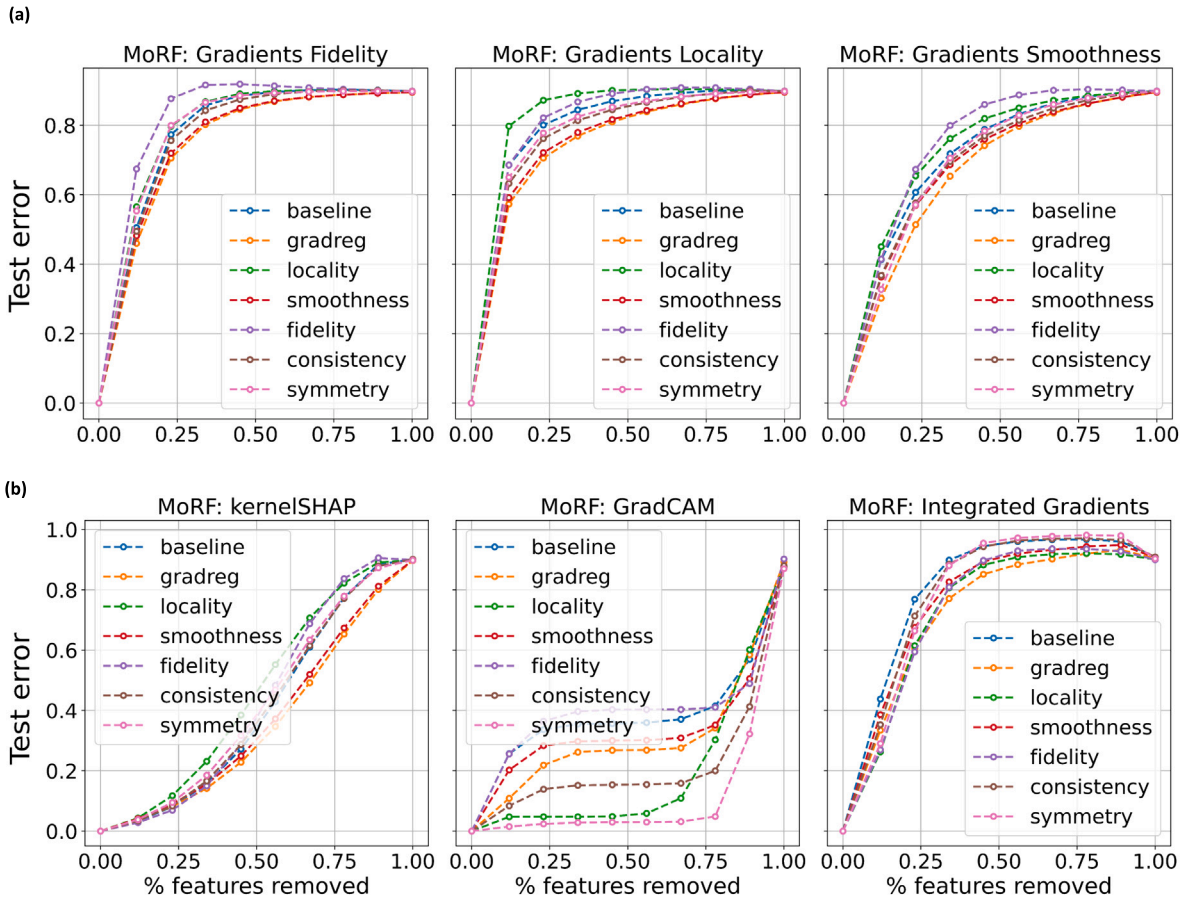

Fig. 4. MoRF curves: (a) test error as a function of the rate of removed features when computing the post-hoc explanation with fidelity (left), locality (middle) or smoothness (right) models; (b) test error as a function of the rate of removed features with explanations computed with kernel SHAP (left), gradcam (middle) or integrated gradients (right).

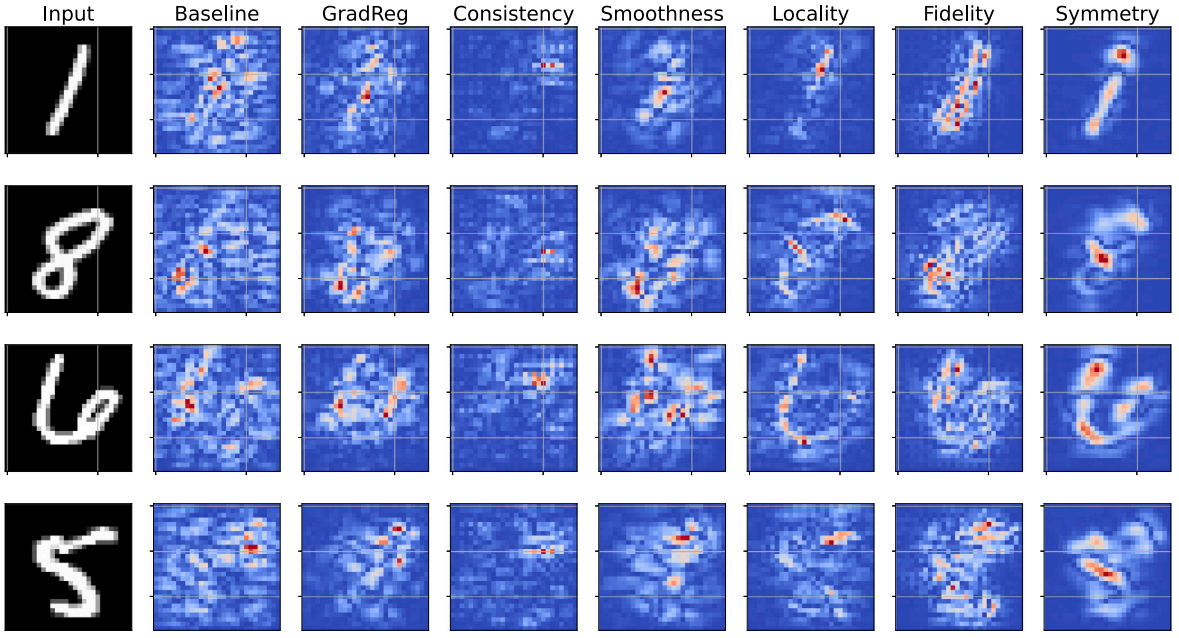

Fig. 5. Saliency maps for some input images. One can notice that explainability-constrained models produce more interpretable attributions with respect to the noisy saliency from the baseline.

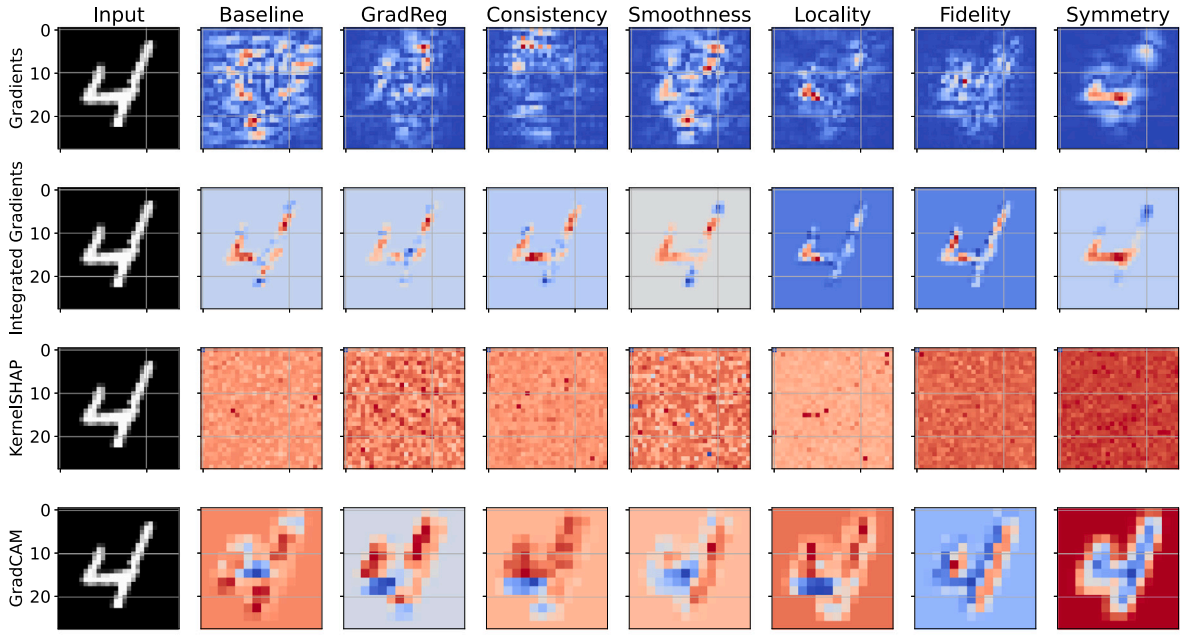

Fig. 6. Saliency maps for some one input image computed with four different xAI methods: Gradients, Integrated Gradients, Kernel SHAP, and GradCAM. Notice the differences between the attributions of the models trained with different constraints for the explanations.

Table 2
Shallow CNNs.

MNIST	Fashion-MNIST
Conv 2D (5×5, 8 kernels)	Conv 2D (3×3, 32 kernels)
Max-pooling (2×2)	Conv 2D (3×3, 64 kernels)
Conv 2D (5×5, 16 kernels)	Max-pooling (2×2)
Max-pooling (2×2)	Dense (32)
Dense (10)	Dense (10)

Table 3
Test accuracy.

Constraint	MNIST	Fashion-MNIST
Baseline	98.7%	92.3%
GradReg	98.5%	90.3%
Locality	98.8%	92.0%
Smoothness	98.5%	89.3%
Fidelity	99.0%	91.6%
Consistency	98.7%	92.3%
Symmetry	98.5%	92.0%

obtain the highest score. It is important to stress that many state-of-the-art xAI methods do not pass the ROAR test [31], while here we obtained better results just with input gradients thanks to the fact that they have been optimized at training time (see also Appendix A.2).

To prove the robustness of the explainability-constrained models against distribution shifts, we inject either spurious square block of random size (uniformly drawn between a minimum and a varying maximum size) and intensity (middle plot in 2b) or Gaussian noise (last plot in 2b) only in the test data without retraining. Here the higher the test accuracy for a given amount of perturbation, the more robust is the model. Ideally, one would like to have a model that retains the same test accuracy regardless of the degree of corruption in the test samples. While a similar behaviour is observed for all models when the perturbation is given by square blocks with increasing size, the smoothness property is crucial for retaining a good test accuracy under Gaussian noise as already shown in [43].

Finally, we make a qualitative comparison between the saliency maps obtained with different attribution priors in Fig. 3. Compared to those obtained from the baseline model (second column in Fig. 3), the gradient-based maps produced with explanation regularization are much more precise and less noisy having non-zero values only in a few localized regions. Regularized models assign zero importance scores to background pixels, which do not contain any information, and this is the case not only for the model trained with locality regularization (8) but also for the other constraints that do not explicitly enforce a notion of locality on the input gradients (see, in particular, the attributions from symmetry in the last column). The optimized feature attributions are also more semantically meaningful and, interestingly, some constraints identify similar patterns or, put in other words, more generalizable explanations (compare e.g. the saliency maps from locality, fidelity and symmetry).

5. Conclusion

With the proliferation of methods for generating explanations, finding ways to select the correct attributions has become increasingly important. We presented a pragmatic framework for optimizing not only the model accuracy but also its interpretability at training time. We compared the saliency maps obtained from models with and without attribution priors and showed that enforcing locality, fidelity and symmetry allowed us to improve significantly the reliability of the final explanations both qualitatively and quantitatively. The framework we discussed is general and flexible enough to allow systematic comparisons between different xAI methods and priors.

The optimization of explanations at training time comes with a computational cost which depends on the specific type of prior, the xAI method used to compute the attributions ϕ at each training step as well as the DL architecture. Nonetheless, it has been shown in [44] that it is possible to reduce the training time by restricting to a class of models for which the attributions can be computed with only a single forward–backward pass, such models need to be non-negatively homogeneous, a property that for most of the known architectures can be achieved by simply removing the bias term of each layer.

Follow-up studies could extend our methodology for comparing the explanations by directly involving human judgement. Following the

Table 4
MoRF curves: AUC scores.

Constraint	Grad locality	Grad fidelity	Grad smoothness	KernelSHAP	GradCAM	IntGrad
Baseline	0.798	0.780	0.711	0.406	0.381	0.811
GradReg	0.747	0.749	0.665	0.355	0.306	0.734
Locality	0.831	0.791	0.732	0.462	0.188	0.736
Smoothness	0.754	0.755	0.685	0.370	0.330	0.770
Fidelity	0.811	0.824	0.749	0.429	0.395	0.744
Consistency	0.778	0.773	0.693	0.411	0.209	0.795
Symmetry	0.784	0.788	0.694	0.422	0.106	0.788

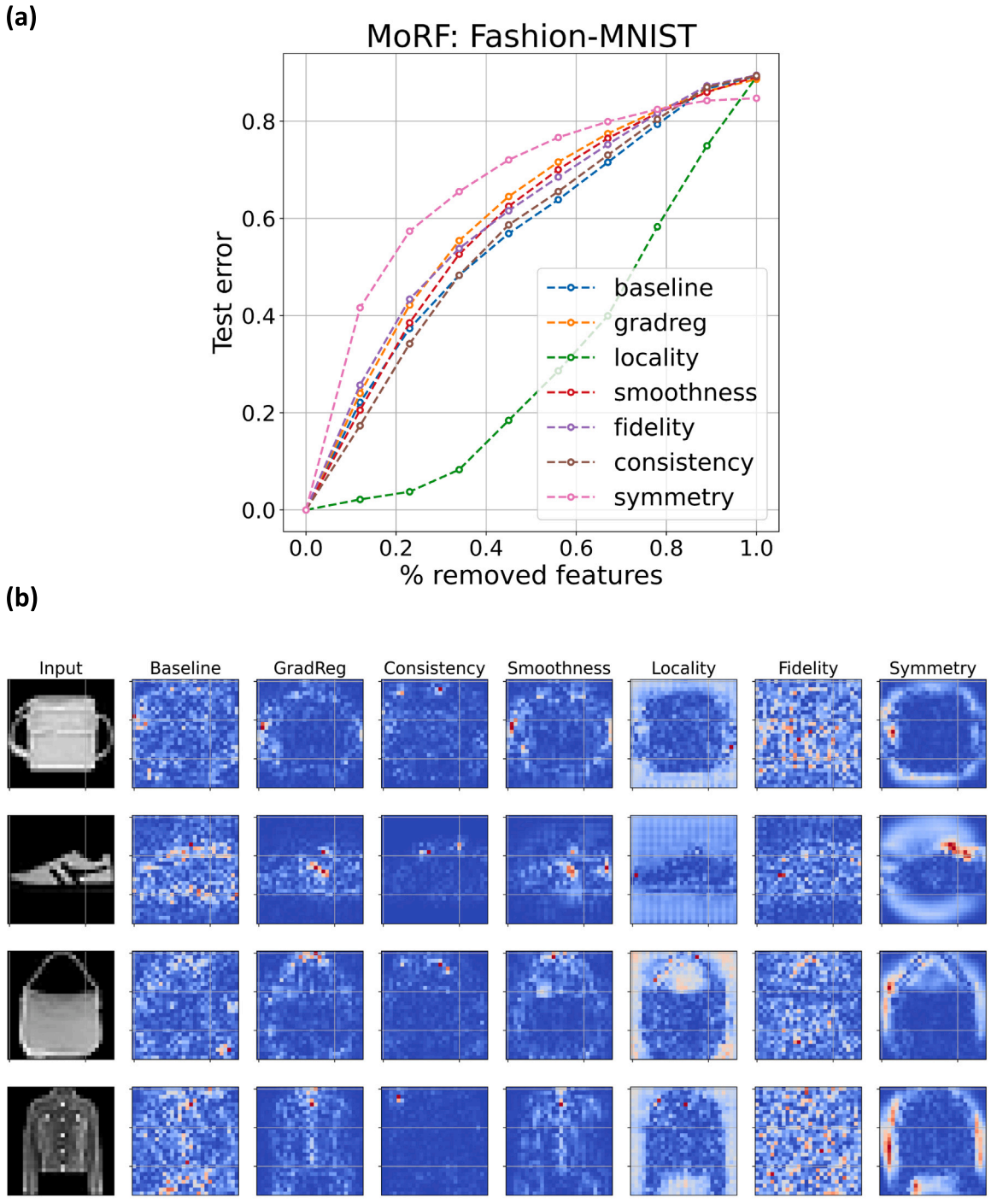

Fig. 7. (a) MoRF curve for all models trained on the Fashion-MNIST dataset. See how the model with symmetry regularization is still able to significantly increase the AUC, while an unsupervised locality penalty deteriorates its performance in this case. b) Fidelity, symmetry and gradient regularization produce more interpretable saliency maps with respect to the noisy saliency from the baseline.

Table 5
Metrics for saliency maps.

Constraint	MoRF	Faithfulness	Complexity
Baseline	0.563	0.744	4.39
GradReg	0.604	**0.837**	4.26
Locality	0.307	0.665	4.41
Smoothness	0.587	0.830	4.35
Fidelity	0.597	0.760	4.40
Consistency	0.626	0.714	**4.25**
Symmetry	**0.664**	0.611	4.35

human-in-the-loop approach (see e.g. [66] and references therein), one could design a user study and ask participants to vote for the most explainable saliency maps for a group of instances. Such an experiment would be costly and time consuming, taking into account that the number of people involved should be quite large in order to reduce as much as possible the variability in responses due to subjective perceptions of interpretability. Nonetheless, when going beyond toy dataset such as MNIST, this kind of human validation would be of crucial importance especially in the application of our analysis to medical datasets where the evaluation even from a small group of experts would be needed to integrate the results from the metrics.

Explainability-constrained training represents a promising direction for solving the conflict between models' performance and the transparency of learned representations. In this regard, theoretical studies are needed to clarify what is the best trade-off and understand model's generalization in terms of both accuracy and explanations. We are confident that our work will encourage further research on novel properties that explanations should hold, which can be learned via regularization in order to advance the disciplines of algorithmic fairness, robustness and trustworthy AI. We foresee applications in fields where domain-knowledge is vast, e.g. health and natural sciences, and where it would be interesting to compare attribution priors with commonly adopted regularization terms such as sparsity or physics conservation laws. In this context, it would also be interesting to combine our approach with the human-in-the-loop framework for what regards both the definition of attribution priors and the evaluation of the produced saliency maps.

CRediT authorship contribution statement

Michele Ronco: Conceptualization, Methodology, Software, Formal analysis, Visualization, Investigation, Writing – original draft. **Gustau Camps-Valls:** Supervision, Writing – review & editing, Funding acquisition.

Declaration of competing interest

The authors declare no competing interests.

Data availability

Data is public and easily accessible through standard DL libraries, code is available at: https://github.com/IPL-UV/xAI-constrained-losses.

Acknowledgments

The authors would like to acknowledge the support from the European Research Council (ERC) under the ERC Synergy Grant USMILE (grant agreement 855187), and the support of the European Union's Horizon 2020 research and innovation program within the project 'DeepCube: Explainable AI pipelines for big Copernicus data' (grant agreement No 101004188).

Appendix

A.1. Model architectures and hyperparameters

Both MNIST and Fashion-MNIST dataset were normalized between 0 and 1. To compare the different attribution priors, we trained shallow CNN architectures reaching nearly state-of-the-art performances on both datasets (about 2% and 5% less accurate than benchmarks respectively) but having relatively short training times depending on the chosen penalty in the loss. All models were developed in PyTorch. Details on the architectures are provided in Table 2. We use $ReLU(x) = max(0, x)$ activation functions and apply batch normalization after each layer. In the Fashion-MNIST CNN we also emply *Dropout* with fraction 0.15 after the convolutional blocks and 0.25 after the first fully connected layer. We tried both 60 and 30 as batch sizes. The optimization is done with Adam where the learning rate is kept equal to 10^{-3}. All experiments are performed with a single 8 GB GPU NVIDIA GeForce RTX 2060.

The regularization parameters λ are obtained through grid search in the range $[0.05, 1.5]$ for each constraint. Overall higher values for λ are used in the Fashion-MNIST. The accuracies on the test set are reported in Table 3.

A.2. Additional results on MNIST

In Fig. 4 we show other MoRF curves obtained over the MNIST dataset. First we show that also fidelity and locality explanations are able to increase the AUC across all models with and without constrained losses. Then we obtain the MoRF by using different xAI methods, i.e.: *Kernel SHAP*, *GradCAM*, and *Integrated Gradients*. Both Kernel SHAP and GradCAM provide less reliable saliency maps according to MoRF. Remarkably, vanilla input-gradient explanations outperform also integrated gradients when penalty terms for explanations are optimized at training time (see again results for symmetry, locality and fidelity in the main text). This suggests that guiding models' explanations is more important that (or at least as important as) picking a specific method for post-hoc explanations. See also Table 4.

More saliency maps for input samples are provided in Fig. 5 where again we can see how explanations obtained with xAI-constrained models are more interpretable also at a qualitative visual level (see Fig. 6).

A.3. Experiments with Fashion-MNIST

Here we report the results obtained with the same regularization terms but training a slightly deeper model (see Table 2) on the Fashion-MNIST dataset. In this case symmetry and fidelity are still able to improve the metrics for the explanations with respect to the unconstrained baseline, while less clear results are obtained with some of the constraints. The scores are reported in Table 5. See Fig. 7.

References

[1] I.J. Goodfellow, Y. Bengio, A.C. Courville, Deep learning, Nature 521 (2015) 436–444.
[2] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT Press, 2016, http://www.deeplearningbook.org.
[3] K. Gade, S.C. Geyik, K. Kenthapadi, V. Mithal, A. Taly, Explainable AI in industry, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 3203–3204.
[4] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, N. Elhadad, 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, 2015.
[5] S.M. Lundberg, B. Nair, M.S. Vavilala, M. Horibe, M.J. Eisses, T. Adams, D.E. Liston, D.K.-W. Low, S.-F. Newman, J. Kim, et al., Explainable machine-learning predictions for the prevention of hypoxaemia during surgery, Nat. Biomed. Eng. 2 (10) (2018) 749.

[6] E. Choi, M.T. Bahadori, J. Sun, J. Kulas, A. Schuetz, W. Stewart, RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism, in: D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, R. Garnett (Eds.), Advances in Neural Information Processing Systems, vol. 29, Curran Associates, Inc., 2016, URL https://proceedings.neurips.cc/paper/2016/file/231141b34c82aa95e48810a9d1b33a79-Paper.pdf.

[7] E. Choi, M.T. Bahadori, A. Schuetz, W.F. Stewart, J. Sun, Doctor AI: Predicting clinical events via recurrent neural networks, in: F. Doshi-Velez, J. Fackler, D. Kale, B. Wallace, J. Wiens (Eds.), Proceedings of the 1st Machine Learning for Healthcare Conference, in: Proceedings of Machine Learning Research, vol. 56, PMLR, Northeastern University, Boston, MA, USA, 2016, pp. 301–318, URL https://proceedings.mlr.press/v56/Choi16.html.

[8] A. Karpatne, G. Atluri, J.H. Faghmous, M.S. Steinbach, A. Banerjee, A.R. Ganguly, S. Shekhar, N.F. Samatova, V. Kumar, Theory-guided data science: A new paradigm for scientific discovery from data, IEEE Trans. Knowl. Data Eng. 29 (2017) 2318–2331.

[9] M. Raissi, P. Perdikaris, G.E. Karniadakis, Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations, J. Comput. Phys. 378 (2019) 686–707.

[10] G. Carleo, I.I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, L. Zdeborov'a, Machine learning and the physical sciences, Rev. Modern Phys. (2019).

[11] P. Holl, N. Thuerey, V. Koltun, Learning to control PDEs with differentiable physics, in: International Conference on Learning Representations, 2020, URL https://openreview.net/forum?id=HyeSin4FPB.

[12] S. Greydanus, M. Dzamba, J. Yosinski, Hamiltonian neural networks, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. dÁlché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, vol. 32, Curran Associates, Inc., 2019, URL https://proceedings.neurips.cc/paper/2019/file/26cd8ecadce0d4efd6cc8a8725cbd1f8-Paper.pdf.

[13] D. Alvarez Melis, T. Jaakkola, Towards robust interpretability with self-explaining neural networks, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems, col. 31, Curran Associates, Inc., 2018, URL https://proceedings.neurips.cc/paper/2018/file/3e9f0fc9b2f89e043bc6233994dfcf76-Paper.pdf.

[14] Z. Chen, Y. Bei, C. Rudin, Concept whitening for interpretable image recognition, Nat. Mach. Intell. 2 (2020) 772–782.

[15] C. Chen, O. Li, D. Tao, A. Barnett, C. Rudin, J.K. Su, This looks like that: Deep learning for interpretable image recognition, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. dÁlché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, vol. 32, Curran Associates, Inc., 2019, URL https://proceedings.neurips.cc/paper/2019/file/adf7ee2dcf142b0e11888e72b43fcb75-Paper.pdf.

[16] G. Montavon, W. Samek, K.-R. Müller, Methods for interpreting and understanding deep neural networks, Digit. Signal Process. 73 (2018) 1–15.

[17] W. Samek, T. Wiegand, K.-R. Müller, Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models, 2017, arxiv, arXiv:1708.08296.

[18] A. Adadi, M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (XAI), IEEE Access 6 (2018) 52138–52160, http://dx.doi.org/10.1109/ACCESS.2018.2870052.

[19] A.B. Arrieta, N.D. Rodríguez, J.D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Inf. Fusion 58 (2020) 82–115.

[20] R. Guidotti, A. Monreale, F. Turini, D. Pedreschi, F. Giannotti, A survey of methods for explaining black box models, ACM Comput. Surv. 51 (2019) 1–42.

[21] G. Vilone, L. Longo, Explainable artificial intelligence: a systematic review, 2020,

[22] M.T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you?": Explaining the predictions of any classifier, in: NAACL 2016, 2016.

[23] S.M. Lundberg, S.-I. Lee, A Unified Approach to Interpreting Model Predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017, pp. 4768–4777.

[24] B. Kim, M. Wattenberg, J. Gilmer, C.J. Cai, J. Wexler, F.B. Viégas, R. Sayres, Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV), in: ICML, 2018.

[25] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, in: Y. Bengio, Y. LeCun (Eds.), 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Workshop Track Proceedings, 2014, URL http://arxiv.org/abs/1312.6034.

[26] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, K.-R. Müller, How to explain individual classification decisions, J. Mach. Learn. Res. 11 (2010) 1803–1831.

[27] L.M. Zintgraf, T.S. Cohen, T. Adel, M. Welling, Visualizing deep neural network decisions: Prediction difference analysis, 2017, CoRR, arXiv:1702.04595, URL http://arxiv.org/abs/1702.04595.

[28] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: D. Precup, Y.W. Teh (Eds.), Proceedings of the 34th International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 70, PMLR, 2017, pp. 3319–3328, URL https://proceedings.mlr.press/v70/sundararajan17a.html.

[29] R.R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: Visual explanations from deep networks via gradient-based localization, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618–626, http://dx.doi.org/10.1109/ICCV.2017.74.

[30] M. Ancona, E. Ceolini, C. Öztireli, M. Gross, Towards better understanding of gradient-based attribution methods for deep neural networks, in: International Conference on Learning Representations, 2018, URL https://openreview.net/forum?id=Sy21R9JAW.

[31] S. Hooker, D. Erhan, P.-J. Kindermans, B. Kim, A benchmark for interpretability methods in deep neural networks, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. dÁlché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, vol. 32, Curran Associates, Inc., 2019, URL https://proceedings.neurips.cc/paper/2019/file/fe4b8556000d0f0cae99daa5c5c5a410-Paper.pdf.

[32] D. Alvarez-Melis, T.S. Jaakkola, On the robustness of interpretability methods, 2018, URL http://arxiv.org/abs/1806.08049, cite arxiv:1806.08049Comment: presented at 2018 ICML Workshop on Human Interpretability in Machine Learning (WHI 2018), Stockholm, Sweden.

[33] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K.T. Schütt, S. Dähne, D. Erhan, B. Kim, The (un)reliability of saliency methods, in: Explainable AI, 2019.

[34] C.-K. Yeh, C.-Y. Hsieh, A. Suggala, D.I. Inouye, P.K. Ravikumar, On the (in)fidelity and sensitivity of explanations, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. dÁlché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, vol. 32, Curran Associates, Inc., 2019, URL https://proceedings.neurips.cc/paper/2019/file/a7471fdc77b3435276507cc8f2dc2569-Paper.pdf.

[35] H. Drucker, Y. Le Cun, Improving generalization performance using double backpropagation, IEEE Trans. Neural Netw. 3 (6) (1992) 991–997, http://dx.doi.org/10.1109/72.165600.

[36] A.S. Ross, M.C. Hughes, F. Doshi-Velez, Right for the right reasons: Training differentiable models by constraining their explanations, in: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17, 2017, pp. 2662–2670, http://dx.doi.org/10.24963/ijcai.2017/371.

[37] A.S. Ros, F. Doshi-Velez, Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients, in: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence, in: AAAI'18/IAAI'18/EAAI'18, AAAI Press, 2018.

[38] L. Rieger, C. Singh, W. Murdoch, B. Yu, Interpretations are useful: Penalizing explanations to align neural networks with prior knowledge, in: H.D. III, A. Singh (Eds.), Proceedings of the 37th International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 119, PMLR, 2020, pp. 8116–8126, URL https://proceedings.mlr.press/v119/rieger20a.html.

[39] A.A. Ismail, H. Corrada Bravo, S. Feizi, Improving deep learning interpretability by saliency guided training, in: M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, J.W. Vaughan (Eds.), Advances in Neural Information Processing Systems, vol. 34, Curran Associates, Inc., 2021, pp. 26726–26739, URL https://proceedings.neurips.cc/paper/2021/file/e0cd3f16f9e883ca91c2a4c24f47b3d9-Paper.pdf.

[40] G. Plumb, M. Al-Shedivat, Á.A. Cabrera, A. Perer, E. Xing, A. Talwalkar, Regularizing black-box models for improved interpretability, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, vol. 33, Curran Associates, Inc., 2020, pp. 10526–10536, URL https://proceedings.neurips.cc/paper/2020/file/770f8e448d07586afbf77bb59f698587-Paper.pdf.

[41] J. Chen, X. Wu, V. Rastogi, Y. Liang, S. Jha, Robust attribution regularization, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. dÁlché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, vol. 32, Curran Associates, Inc., 2019, URL https://proceedings.neurips.cc/paper/2019/file/172ef5a94b4dd0aa120c6878fc29f70c-Paper.pdf.

[42] V. Pillai, H. Pirsiavash, Explainable models with consistent interpretations, Proc. AAAI Conf. Artif. Intell. 35 (3) (2021) 2431–2439, URL https://ojs.aaai.org/index.php/AAAI/article/view/16344.

[43] G. Erion, J.D. Janizek, P. Sturmfels, S.M. Lundberg, S.-I. Lee, Improving performance of deep learning models with axiomatic attribution priors and expected gradients, Nat. Mach. Intell. (2021) 1–12.

[44] R. Hesse, S. Schaub-Meyer, S. Roth, Fast axiomatic attribution for neural networks, in: M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, J.W. Vaughan (Eds.), Advances in Neural Information Processing Systems, vol. 34, Curran

Associates, Inc., 2021, pp. 19513–19524, URL https://proceedings.neurips.cc/paper/2021/file/a284df1155ec3e67286080500df36a9a-Paper.pdf.

[45] T. Cohen, M. Welling, Group equivariant convolutional networks, in: M.F. Balcan, K.Q. Weinberger (Eds.), Proceedings of the 33rd International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 48, PMLR, New York, New York, USA, 2016, pp. 2990–2999, URL https://proceedings.mlr.press/v48/cohenc16.html.

[46] K. Lenc, A. Vedaldi, Understanding image representations by measuring their equivariance and equivalence, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 991–999.

[47] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, K.R. Klaus Robert Müller, Evaluating the visualization of what a deep neural network has learned, IEEE Trans. Neural Netw. Learn. Syst. 28 (2017) 2660–2673, http://dx.doi.org/10.1109/TNNLS.2016.2599820.

[48] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, B. Kim, Sanity checks for saliency maps, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems, vol. 31, Curran Associates, Inc., 2018, URL https://proceedings.neurips.cc/paper/2018/file/294a8ed24b1ad22ec2e7efea049b8737-Paper.pdf.

[49] R.J. Tomsett, D. Harborne, S. Chakraborty, P.K. Gurram, A.D. Preece, Sanity checks for saliency metrics, in: AAAI, 2020.

[50] L. Rieger, L.K. Hansen, IROF: a low resource evaluation metric for explanation methods, 2020, CoRR, arXiv:2003.08747, URL https://arxiv.org/abs/2003.08747.

[51] U. Bhatt, A. Weller, J.M.F. Moura, Evaluating and aggregating feature-based model explanations, in: C. Bessiere (Ed.), Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20, International Joint Conferences on Artificial Intelligence Organization, 2020, pp. 3016–3022, http://dx.doi.org/10.24963/ijcai.2020/417.

[52] H. Shah, P. Jain, P. Netrapalli, Do input gradients highlight discriminative features? in: M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, J.W. Vaughan (Eds.), Advances in Neural Information Processing Systems, vol. 34, Curran Associates, Inc., 2021, pp. 2046–2059, URL https://proceedings.neurips.cc/paper/2021/file/0fe6a94848e5c68a54010b61b3e94b0e-Paper.pdf.

[53] A. Ansuini, A. Laio, J.H. Macke, D. Zoccolan, Intrinsic dimension of data representations in deep neural networks, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. dÁlché-Buc, E. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems, vol. 32, Curran Associates, Inc., 2019, URL https://proceedings.neurips.cc/paper/2019/file/cfcce0621b49c983991ead4c3d4d3b6b-Paper.pdf.

[54] B. Recht, R. Roelofs, L. Schmidt, V. Shankar, Do ImageNet classifiers generalize to ImageNet? 2019, arxiv, arXiv:1902.10811.

[55] I. Lage, A.S. Ross, B. Kim, S.J. Gershman, F. Doshi-Velez, Human-in-the-loop interpretability prior, Adv. Neural Inf. Process. Syst. 31 (2018).

[56] T. Fel, D. Vigouroux, R. Cadène, T. Serre, How Good is your Explanation? Algorithmic Stability Measures to Assess the Quality of Explanations for Deep Neural Networks, in: 2022 CVF Winter Conference on Applications of Computer Vision (WACV), Hawaii, United States, 2022, URL https://hal.archives-ouvertes.fr/hal-02930949.

[57] O. Bousquet, A. Elisseeff, Stability and generalization, J. Mach. Learn. Res. 2 (2002) 499–526, http://dx.doi.org/10.1162/153244302760200704.

[58] C. Shorten, T.M. Khoshgoftaar, A survey on image data augmentation for deep learning, J. Big Data 6 (2019) 1–48.

[59] A. Mikołajczyk, M. Grochowski, Data augmentation for improving deep learning in image classification problem, in: 2018 International Interdisciplinary PhD Workshop (IIPhDW), 2018, pp. 117–122, http://dx.doi.org/10.1109/IIPHDW.2018.8388338.

[60] L. Perez, J. Wang, The effectiveness of data augmentation in image classification using deep learning, 2017, arxiv, arXiv:1712.04621.

[61] T. Cohen, M. Weiler, B. Kicanaoglu, M. Welling, Gauge equivariant convolutional networks and the icosahedral CNN, in: ICML, 2019.

[62] M. Bronstein, J. Bruna, Y. Lecun, A. Szlam, P. Vandergheynst, Geometric deep learning: Going beyond euclidean data, IEEE Audio Electroacoust Newslett 34 (4) (2017) 18–42, http://dx.doi.org/10.1109/MSP.2017.2693418, Publisher Copyright: © 2016 IEEE..

[63] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2015, CoRR, arXiv:1412.6980.

[64] Y. LeCun, C. Cortes, MNIST handwritten digit database, 2010, URL http://yann.lecun.com/exdb/mnist/.

[65] L. Arras, G. Montavon, K.-R. Müller, W. Samek, Explaining recurrent neural network predictions in sentiment analysis, in: Proceedings of the 8th Workshop on Computational Approaches To Subjectivity, Sentiment and Social Media Analysis, Association for Computational Linguistics, 2017, pp. 159–168, http://dx.doi.org/10.18653/v1/W17-5221, URL https://aclanthology.org/W17-5221.

[66] A. Holzinger, Interactive machine learning for health informatics: when do we need the human-in-the-loop? Brain Inf. 3, 119–131, http://dx.doi.org/10.1007/s40708-016-0042-6.

Michele Ronco (born 1991 in Rome) studied physics at the University of Rome "La Sapienza", where he obtained his Ph.D. in 2019 with a thesis on the phenomenology of quantum gravity approaches in astrophysical observations. During the Ph.D. he was a visitor scientist in several research institutions, among which Penn State University, Fudan University, IEM-CSIC, and University of Valencia. He then joined the HESS telescope group at the University Pierre and Marie Curie in Paris where he worked on tests and simulations of Lorentz invariance violation models in the propagation of high-energy gamma rays. After that, he worked as data scientist in the industry both in the insurance and in the remote sensing sectors, and applied CNN models for the classification and segmentation of Sentinel-2 images. He is currently a Postdoctoral Researcher at the Image Processing Laboratory of the University of Valencia with a focus on explainable machine learning methods for a variety of earth science problems, ranging from wildfire forecasting to human movements induced by weather hazards. He presented his works at several international conferences, workshops and invited seminars, and published over 20+ peer-reviewed international journal papers.

Gustau Camps-Valls (born 1972 in Valncia) is a Physicist and Full Professor in Electrical Engineering in the Universitat de Valéncia, Spain, where lectures on machine learning, remote sensing and signal processing. He is the Head of the Image and Signal Processing (ISP) group, an interdisciplinary group of 40 researchers working at the intersection of AI for Earth and Climate sciences. Prof. Camps-Valls published over 250+ peer-reviewed international journal papers, 350+ international conference papers, 25 book chapters, and 5 international books on remote sensing, image processing and machine learning. He has an h-index of 82 with 31000+ citations in Google Scholar. He was listed as a Highly Cited Researcher in 2011, 2021 and 2022; currently has 13 ≪Highly Cited Papers≫ and 1 ≪Hot Paper≫, Thomson Reuters ScienceWatch identified his activities as a Fast Moving Front research (2011) and the most-cited paper in the area of Engineering in 2011, received the Google Classic paper award (2019), and Stanford Metrics includes him in the top 2 percent most cited researchers of 2017-2020. He publishes in both technical and scientific journals, from IEEE and PLOS One to Nature, Nature Communications, Science Advances, and PNAS. He has been Program Committee member of international conferences (IEEE, SPIE, EGU, AGU), and Technical Program Chair at IEEE IGARSS 2018 (2400+ attendees) and general at AISTATS 2022. He served in technical committees of the IEEE GRSS and IEEE SPS, as Associate Editor of 5 top IEEE journals, and in the prestigious IEEE Distinguished Lecturer program of the GRSS (2017-2019) to promote ≪AI in Earth sciences≫ globally. He has given 100+ talks, keynote speaker in 10+ conferences, and (co)advised 10+ PhD theses. He coordinated/participated in 60+ research projects, involving industry and academia at national and European levels. He assisted the aerospace industry in Advisory Boards; Fellow Consultant of the ESA PhiLab (2019) and member of the EUMETSAT MTG-IRS Science Team. He is compromised with open source/access in Science, and is habitual panel evaluator for H2020 (ERC, FET), NSF, China and Swiss Science Foundations. He coordinates the 'Machine Learning for Earth and Climate Sciences' research program of ELLIS, the top network of excellence on AI in Europe. He was elevated to IEEE Fellow member (2018) in two Societies (Geosciences and Signal Processing) and to ELLIS Fellow (2019). Prof. Camps-Valls is the only researcher receiving two European Research Council (ERC) grants in two different areas: an ERC Consolidator (2015, Computer Science) and ERC Synergy (2019, Physical Sciences) grants to advance AI for Earth and Climate Sciences. In 2021 he became a Member of the ESSC panel part of the European Science Foundation (ESF), and in 2022 was elevated to Fellow of the European Academy of Sciences (EurASc), Fellow of the Academia Europeae (AE), and Fellow of Asia-Pacific Artificial Intelligence Association (AAIA).