This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Structured Gradient-based Interpretations via Norm-Regularized Adversarial Training

Shizhan Gong, Qi Dou, Farzan Farnia
The Chinese University of Hong Kong
{szgong22, qdou, farnia}@cse.cuhk.edu.hk

Abstract

Gradient-based saliency maps have been widely used to explain the decisions of deep neural network classifiers. However, standard gradient-based interpretation maps, including the simple gradient and integrated gradient algorithms, often lack desired structures such as sparsity and connectedness in their application to real-world computer vision models. A frequently used approach to inducing sparsity structures into gradient-based saliency maps is to alter the simple gradient scheme using sparsification or norm-based regularization. A drawback with such post-processing methods is their frequently-observed significant loss in fidelity to the original simple gradient map. In this work, we propose to apply adversarial training as an in-processing scheme to train neural networks with structured simple gradient maps. We show a duality relation between the regularized norms of the adversarial perturbations and gradient-based maps, based on which we design adversarial training loss functions promoting sparsity and group-sparsity properties in simple gradient maps. We present several numerical results to show the influence of our proposed norm-based adversarial training methods on the standard gradient-based maps of standard neural network architectures on benchmark image datasets[1].

1. Introduction

Deep neural networks have attained remarkable results in various computer vision tasks including image classification [17], object detection [44], and semantic segmentation [10]. However, understanding and explaining the decision-making process of these complex models remains a challenge, which is required for high-risk applications such as medical imaging [27], autonomous driving [16], and face recognition [37]. To interpret the predictions of neural network classifiers, several explanation methodologies have been proposed in the literature. Among these explanation methods, *gradient-based saliency maps* have been frequently applied to interpret image classifiers. The saliency maps provide insight into the regions of the input image contributing to the classifier's decision and potentially prompt the empirical understanding of phenomena from data-driven models.

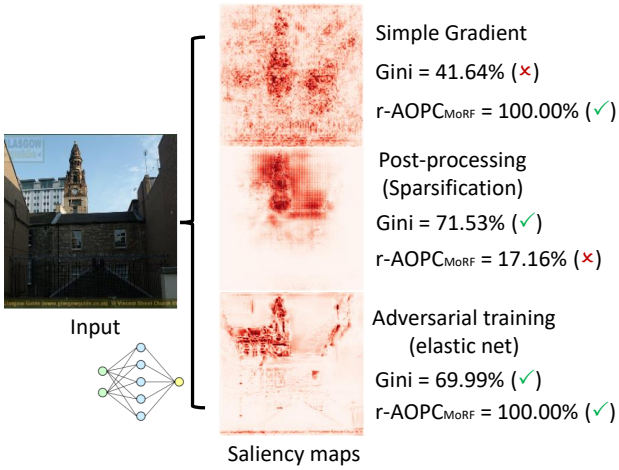

Figure 1. The original simple gradient map could be dense and noisy. Post-processing methods such as sparsification enhance the sparsity at the expense of lower fidelity to the original map. Our proposed in-processing strategy with adversarial training results in higher sparsity without losing fidelity to the simple-grad map. (We use the Gini index as a sparsity measure, and relative AOPC$_{MoRF}$ with respect to simple gradient as a fidelity score.)

Despite the widespread use of standard gradient-based interpretation maps such as simple gradients [29] and integrated gradients [32], they often lack desired structures such as sparsity and connectedness. When applied to real-world computer vision tasks, the generated saliency maps have been frequently observed to be noisy and dense. Such unstructured behavior could hinder the application of gradient-based maps for detecting the influential features in the classifier's decision-making process. To obtain more structured saliency maps, several post-processing

[1]The paper's code is available at: https://github.com/peterant330/AdvGrad

schemes [18, 30, 31, 42, 43] have been proposed in the literature, which denoise the saliency map to gain higher sparsity. However, a drawback of these post-processing methods is the deviation from the original simple gradient map that alters the definition of the interpretation map. For example, Adebayo et al. [1] conducted a sanity check and found several interpretation methods such as Guided Propagation and Integrated Gradients lacked sensitivity to the model or the data-generating process. Arun et al. [2] demonstrated similar findings on medical imaging applications. To illustrate the fidelity vs. sparsity trade-off in such post-processing schemes, we display an example in Fig. 1, where the simple gradient map looks to contain significant noise scattered over the background. A post-processing sparsification of the gradient map could raise the sparsity of the map; however, the sparsity comes at the cost of a considerable drop (more than 82%) in the fidelity score.

In this work, we propose a unified adversarial training (AT) framework to address the unstructured nature of standard simple-gradient maps, where we apply a norm-regularized adversarial training with a properly-designed norm function as an in-processing scheme to promote the sparsity structures in the simple gradient map. AT with standard L_2 and L_∞-norm perturbations [19] has been successfully and widely utilized in the literature to enhance the adversarial robustness [33] and interpretability [26] of deep neural networks. In our analysis, we extend the understanding of the influence of AT algorithms on the interpretability of a trained model, where we provide a convex duality framework to understand the influence of the norm function constrained in AT on the regularized norm of gradient maps and hence the interpretability of neural nets.

Specifically, we show a duality relation between the regularized norm of adversarial perturbations and the norm of input-based gradients. This duality relationship allows us to design norm-regularized adversarial training methods, which translates into the regularization of standard sparsity-inducing norms, e.g. the group norm [41] and the elastic net [45], of the simple gradient maps. As we show in this work, the special cases of our proposed framework can promote desired characteristics such as sparsity and connectedness in simple gradient maps. Importantly, unlike the sparsity-regularizing post-processing schemes, the gained sparsity properties of our proposed AT methods do not cost any loss in fidelity to the simple-gradient map, because we do not alter the definition of the interpretation map and only regularize the training process toward networks with more structured gradient maps.

We conduct several numerical experiments on benchmark image datasets to validate the efficacy of our proposed AT-based methodology. The numerical results reveal the impact of our proposed norm-based AT methods on standard gradient-based maps, showing the enhanced sparsity and connectedness properties. We empirically analyze the performance of the trained neural nets in terms of several factors including interpretability, robustness, and stability. Our numerical results indicate the improvements in the mentioned structure-related factors offered by our designed norm-regularized AT methods. Finally, we utilize the shown duality relation to propose an interpretation harmonization scheme for aligning simple gradient maps with expert gaze maps, which performs satisfactorily in our numerical experiments. Our work's main contributions can be summarized as:

- We show a duality relation between the regularized norms of adversarial perturbations and gradient maps, and develop a unified AT framework for regularizing gradient maps.
- We derive special cases of our proposed AT-based framework leading to the regularization of the elastic net and group-norm of gradient map.
- We leverage the duality framework to propose an interpretation harmonization scheme for aligning gradient maps with human attention.
- We provide numerical results showing the efficacy of the AT-based framework in improving the sparsity and stability of interpretation maps.

2. Related Work

Gradient-based Interpretation. Using the gradient of the output of a deep neural network w.r.t. an input image is a widely-used approach to generate saliency maps [29]. This method has been utilized in several related works and multiple variants of this approach are proposed in the literature, including SmoothGrad [30], Integrated Gradients [32], DeepLIFT [28], and GradCAM [24]. The vanilla gradient-based maps without post-processing are often noisy and difficult to interpret. A group of methods are proposed to improve the visual quality by sparsifying the gradient-based maps: Guided Propagation [31] removes the noise artifacts from saliency maps by suppressing negative activations in the back-propagation. Sparsified-SmoothGrad [18] applies a sparsification to the saliency map to promote sparsity and robustness. MoreauGrad [43] generates saliency maps using an optimization problem with sparsity regularization. One drawback of the discussed methods is that they are all post-processing schemes, which may compromise the fidelity to the original simple-gradient map. On the other hand, our proposed framework is an in-processing scheme that remains faithful to the simple gradient map.

Adversarial Training and Interpretability. Adversarial training (AT) involves training the networks using adversarial examples [19]. Fast Gradient Sign Method (FGSM) [9] generates adversarial samples using a single iteration to create the adversarial training example for training the classifier. Projected gradient descent (PGD) [19] uses several

gradient-based steps to generate more effective adversarial examples in order to make the trained model more robust to norm-constrained adversarial attacks. While the primary objective of AT is to improve the trained model's robustness against adversarial attacks, several studies have shown that AT can also improve the visual quality of interpretation maps [15, 22, 26]. Specifically, AT has been empirically shown to suppress irrelevant features [14], and the standard ℓ_∞-norm-based AT will lead to higher sparsity and stability of the interpretation maps [3]. Building upon and extending these findings, our work provides a unified duality framework that connects the perturbation norm constrained by the AT algorithm and the norm penalized for the interpretation map. Specifically, we discuss novel variants of AT methods that lead to the regularization of the elastic net and group norms of the interpretation maps which have not been studied in these related works.

3. Preliminaries

3.1. Gradient-based Saliency Maps

Throughout this work, our analysis mainly focuses on the standard simple gradient as the gradient-based interpretation. We use $\mathbf{x} \in \mathbb{R}^d$ and $\mathbf{y} \in \mathbb{R}^C$ to denote the input and output of the neural network classifier. Note that $\mathbf{y}$ denotes the output of the neural net's post-softmax layer and contains the assigned likelihood for each of the C classes in the classification task. The neural net function $f_\theta : \mathbb{R}^d \to \mathbb{R}^C$ maps the input vector to the output vector, where θ denotes the parameters of the neural net. The simple gradient generates a saliency map by taking derivatives of the output of the neural network with respect to the input:

$$\mathrm{SG}(f_{\theta,c}, \mathbf{x}) := \nabla_{\mathbf{x}} f_{\theta,c}(\mathbf{x}).$$

Here, c can be chosen as ground-truth label y or the label c with the maximum likelihood assigned by the classifier. In our analysis, we assume $\nabla_{\mathbf{x}} f_\theta(\mathbf{x})$ denotes the saliency map corresponding to the ground-truth label of sample $\mathbf{x}$.

3.2. FGSM and PGD

We consider two of the most widely used forms of adversarial training, FGSM and PGD, in this work. These methods first generate adversarial samples by calculating the gradient of the loss function with respect to the input data and then perturb the data by taking one or multiple small steps in the direction of the sign of the gradient. The perturbation is usually bounded by l_∞-norm. The parameters of the network are then updated to minimize the loss with respect to these adversarial samples. We can use the following objective to summarize the optimization problem where $\widehat{P}$ is the empirical distribution of training samples

$$\min_\theta \ \mathbb{E}_{(\mathbf{x},y)\sim\widehat{P}}\Big[\max_{\|\delta\|_\infty \le \epsilon} \ \mathcal{L}(f_\theta(\mathbf{x}+\delta), y)\Big]. \qquad (1)$$

Table 1. Summary of regularized norm h on perturbation δ and the resulting regularized function $h^\star$ of the input gradient $\nabla_{\mathbf{x}}\mathcal{L}(f_\theta(\mathbf{x}), y)$. Detailed description can be found in Section 4.3.

$h(\delta)$	$h^\star(\nabla_{\mathbf{x}}\mathcal{L}(f_\theta(\mathbf{x}), y))$
$\mathbb{I}(\|\delta\|_\infty \le \epsilon)$	$\epsilon\|\nabla_{\mathbf{x}}\mathcal{L}(f_\theta(\mathbf{x}), y)\|_1$
$\mathbb{I}(\|\delta\|_{2,\infty} \le \epsilon)$	$\epsilon\|\nabla_{\mathbf{x}}\mathcal{L}(f_\theta(\mathbf{x}), y)\|_{2,1}$
$\sum_i PQ_{\epsilon_1,\epsilon_2}(\delta_i)$	$\epsilon_1\|\nabla_{\mathbf{x}}\mathcal{L}(f_\theta(\mathbf{x}), y)\|_1 + \epsilon_2\|\nabla_{\mathbf{x}}\mathcal{L}(f_\theta(\mathbf{x}), y)\|_2^2$
$\frac{1}{4\epsilon}\|\frac{1}{\mathbf{W}} \odot \delta\|_2^2$	$\epsilon\|\mathbf{W} \odot \nabla_{\mathbf{x}}\mathcal{L}(f_\theta(\mathbf{x}), y)\|_2^2$

FGSM generates adversarial samples by taking one step in the direction of the sign of the gradient:

$$\mathbf{x}^*_{\mathrm{FGSM}} = \mathbf{x} + \epsilon \cdot \mathrm{sign}\big(\nabla_{\mathbf{x}}\mathcal{L}(f_\theta(\mathbf{x}), y)\big)$$

PGD, on the other hand, iteratively updates the samples with a stepsize α:

$$\mathbf{x}_{t+1} = \Pi_{\mathbf{x}+S}\big(\mathbf{x}_t + \alpha\mathrm{sign}(\nabla_{\mathbf{x}_t}\mathcal{L}(f_\theta(\mathbf{x}_t), y))\big),$$

where $S = \{\mathbf{z} : \|\mathbf{z}\|_\infty \le \epsilon\}$ denotes the ϵ-radius L_∞-ball.

3.3. Fenchel Conjugate

The Fenchel conjugate is an essential operation in convex analysis which has many applications to various computer vision algorithms. For a function $f : \mathcal{X} \to \mathbb{R}$, its Fenchel conjugate $f^\star : \mathcal{X} \to \mathbb{R}$ is defined as:

$$f^\star(z) = \sup\big\{\langle x, z\rangle - f(x) : x \in \mathcal{X}\big\},$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product, and sup represents the supremum. Note that for a convex f, f will be the Fenchel conjugate of $f^\star$. Furthermore, for every function f, the Fenchel conjugate $f^\star$ is guaranteed to be convex.

4. An Adversarial Training Methodology for Regularizing Interpretation Maps

In this section, we show a duality relation between the input gradient and the perturbation vectors, which is based on the linear approximation of the loss function $\ell\big(f(\mathbf{x}+\delta), y\big)$ around data point $\mathbf{x}$. In the approximation, the first-order term $\langle\nabla_{\mathbf{x}}\ell\big(f(\mathbf{x}), y\big), \delta\rangle$ is the inner product of the perturbation vector and the input-based gradient of the loss function. Therefore, as will be explained in the section, to penalize a norm of the saliency map in the adversarial training process, we propose to constrain the dual norm of the perturbation vector. We summarize three variants of the norm constraint on the perturbation and its effect on the simple gradient saliency maps in Table 1.

4.1. Rethinking the Impact of ℓ_∞-Norm-Based Adversarial Training on Saliency Maps

We perform a first-order analysis by means of Fenchel conjugate to indicate how L_∞-norm-based adversarial training

could lead to higher sparsity in gradient-based maps. Under a small enough perturbation norm bound ϵ and twice-differentiable loss function, the objective of adversarial training in Eq. 1 can be approximated by first-order Taylor expansion as

$$\text{Eq. 1} = \min_{\theta} \mathbb{E}_{\mathbf{x},y\sim\widehat{P}}\Big[\mathcal{L}(f_\theta(\mathbf{x}),y) + \max_{\|\delta\|_\infty\leq\epsilon}\left\{\delta^T\nabla_{\mathbf{x}}\mathcal{L}(f_\theta(\mathbf{x}),y)\right\}\Big] + \mathcal{O}(\epsilon^2).$$

Using the common cross-entropy loss or hinge loss, the derivative of the loss with respect to the predicted logits at every training sample will be $\nabla_{\mathbf{x}}\mathcal{L}(f_\theta,(\mathbf{x}),y) = -f_{\theta,y}(\mathbf{x})\nabla_{\mathbf{x}}f_\theta(\mathbf{x})$, so we can build the connection between the objective and the saliency map:

$$\text{Eq. 1} = \min_{\theta} \mathbb{E}_{\mathbf{x},y\sim\widehat{P}}\Big[\mathcal{L}(f_{\theta,y}(\mathbf{x})) + f_{\theta,y}(\mathbf{x})\max_{\|\delta\|_\infty\leq\epsilon}\left\{-\delta^T\nabla_{\mathbf{x}}f_\theta(\mathbf{x})\right\}\Big] + \mathcal{O}(\epsilon^2).$$

Note we can leverage Fenchel conjugate operation to transform the maximization subproblem in the above problem into an unconstrained optimization:

$$\max_{\|\delta\|_\infty\leq\epsilon} -\delta^T\nabla_{\mathbf{x}}f_\theta(\mathbf{x}) = \max_{\delta} -\delta^T\nabla_{\mathbf{x}}f_\theta(\mathbf{x}) - \mathbb{I}(\|\delta\|_\infty\leq\epsilon), \tag{2}$$

where $\mathbb{I}(\text{condition}) = 0$ if the condition holds and $+\infty$ otherwise. With the concept of Fenchel conjugate, the optimal value of the optimization problem defined in Eq. 2 is $\epsilon\|\nabla_{\mathbf{x}}f_\theta(\mathbf{x})\|_1$. Hence, we successfully transform the original minimax optimization problem into a regularized optimization within an approximation error $\mathcal{O}(\epsilon^2)$:

$$\text{Eq. 1} \approx \min_{\theta} \mathbb{E}_{\mathbf{x},y\sim\widehat{P}}\Big[\mathcal{L}(f_\theta(\mathbf{x}),y) + \epsilon f_{\theta,y}(\mathbf{x})\|\nabla_{\mathbf{x}}f_\theta(\mathbf{x})\|_1\Big].$$

Hence, we find that adversarial training with L_∞-norm is approximately equivalent to applying L_1-norm regularization to the simple gradient saliency map, therefore promoting the sparsity of the saliency maps.

4.2. Adversarial Training with Fenchel Conjugate Regularization: a Unified Approach

Next, we consider adversarial training with a general norm-based regularization penalty. Here our goal is to promote a structured saliency map through a regularization function $h^\star(\mathbf{z})$. Note that for a convex h, the Fenchel conjugate of $h^\star(\mathbf{z})$ will be $h(\mathbf{x})$, Therefore, we consider the min-max optimization with the adversarial penalty function $-h(\delta)$:

$$\min_{\theta} \mathbb{E}_{\mathbf{x},y\sim\widehat{P}}\Big[\max_{\delta}\ \mathcal{L}(f_\theta(\mathbf{x}+\delta),y) - h(\delta)\Big].$$

Following our analysis for the L_∞-norm-based adversarial training, we define the first-order approximate loss:

$$\widehat{\mathcal{L}}(f_\theta(\mathbf{x}),y,\delta) := \mathcal{L}(f_\theta(\mathbf{x}),y) + \delta^T\nabla_x\mathcal{L}(f_\theta(\mathbf{x}),y),$$

To measure how well $\widehat{\mathcal{L}}(f_\theta(\mathbf{x}),y,\delta)$ approximates the adversarial loss, we observe the approximation error can be simply bounded for the class of λ-smooth functions.

Definition 1 *We call $g:\mathbb{R}^d\to\mathbb{R}$ λ-smooth if for every $\mathbf{x}$ and $\mathbf{z}$: $\|\nabla g(\mathbf{x}) - \nabla g(\mathbf{z})\|_2 \leq \lambda\|\mathbf{x}-\mathbf{z}\|_2$.*

Observation 1 *The approximation error of the first-order Taylor series expansion of adversarial loss under an ϵ-L_2-norm bounded perturbation $\|\delta\|\leq\epsilon$ can be bounded as*

$$\left|\widehat{\mathcal{L}}(f_\theta(\mathbf{x}),y,\delta) - \mathcal{L}(f_\theta(\mathbf{x}+\delta),y)\right| \leq \frac{\lambda\epsilon^2}{2}$$

The observation shows $\widehat{\mathcal{L}}(f_\theta(\mathbf{x}),y,\delta)$ can be a good approximation of the adversarial loss under a small $\|\delta\|$. Combining the approximate loss with $h(\delta)$ regularization, we can derive the following equivalence by Fenchel conjugate:

Proposition 1 *Using Fenchel conjugate $h^\star$, we can reduce the maximization of the approximate adversarial loss as*

$$\begin{aligned}&\max_{\delta}\left\{\widehat{\mathcal{L}}(f_\theta(\mathbf{x}),y,\delta) - h(\delta)\right\}\\ &= \mathcal{L}(f_\theta(\mathbf{x}),y) + h^\star(\nabla_{\mathbf{x}}\mathcal{L}\big(f_\theta(\mathbf{x}),y\big)\big).\end{aligned} \tag{3}$$

In a scenario that the neural network completely fits the training data, we have that $\nabla_{\mathbf{x}}\mathcal{L}(f_\theta(\mathbf{x}),y) = -f_{\theta,y}(\mathbf{x})\nabla_x f_\theta(\mathbf{x}) \approx -\nabla_x f_\theta(\mathbf{x})$. Therefore, to enforce the gradient map $\nabla_x f_\theta(\mathbf{x})$ to satisfy a bounded penalty function $h^\star$, we can conduct adversarial training with an additive penalty term $-h(\delta)$. The optimal perturbation δ^* can be derived via the standard gradient ascent algorithm to maximize the regularized adversarial loss or be approximated with the analytic solution to the approximate optimization problem.

4.3. Special Cases of the Fenchel Conjugate-based Duality Framework

Leveraging the duality framework in the previous section, we derive regularization penalty terms to penalize other sparsity norms different from l_1-norm and discuss their influence on the saliency maps.

$L_{2,1}$-group-norm penalty. The $L_{2,1}$-group-norm of $\mathbf{x}\in\mathbb{R}^d$ for disjoint variable subsets $S_1,\cdots,S_t\subseteq\{1,\cdots,d\}$ is defined as the special case $p=1$ of the following definition:

$$\|\mathbf{x}\|_{2,p} := \|[\|\mathbf{x}_{S_1}\|_2,\cdots,\|\mathbf{x}_{S_t}\|_2]\|_p.$$

To apply the duality framework to the group norm (with coefficient $\epsilon>0$), we derive its Fenchel conjugate as follows.

Proposition 2 *The Fenchel conjugate of the above $L_{2,1}$-group-norm is $h(\mathbf{z}) = \mathbb{I}(\|\mathbf{z}\|_{2,\infty}\leq\epsilon)$.*

Different from L_1-norm inducing pixel-level sparsity, $L_{2,1}$-group-norm will induce sparsity of the patch basis. Under this regularization, the more influential patches are expected to be highlighted in the saliency map.This effect could further lead to a connected saliency map, which will be desired when adjacent pixels belonging have similar effects on the prediction of the classifier.

Elastic net penalty. The elastic net penalty $h^\star(\mathbf{x})$ linearly combines the L_1 and L_2^2 penalty terms:

$$h^\star(\mathbf{x}) = \epsilon_1\|\mathbf{x}\|_1 + \epsilon_2\|\mathbf{x}\|_2^2, \epsilon_1 > 0, \epsilon_2 > 0.$$

Proposition 3 *The Fenchel conjugate of the above elastic net penalty $h^\star$ is $h(\mathbf{z}) = \sum_i PQ_{\epsilon_1,\epsilon_2}(\mathbf{z}_i)$ where $PQ_{\epsilon_1,\epsilon_2}$ is the piece-wise quadratic function defined as:*

$$PQ_{\epsilon_1,\epsilon_2}(\mathbf{z}_i) = \begin{cases} \frac{1}{4\epsilon_2}(\mathbf{z}_i - \epsilon_1)^2 & \text{if } \epsilon_1 < \mathbf{z}_i \\ 0 & \text{if } -\epsilon_1 \le \mathbf{z}_i \le \epsilon_1 \\ \frac{1}{4\epsilon_2}(\mathbf{z}_i + \epsilon_1)^2 & \text{if } \mathbf{z}_i < -\epsilon_1. \end{cases}$$

The linear combination of the L_1-norm and L_2^2-norm penalty terms in the elastic net results in a strongly convex regularization function, which is known to promote the stability of the sparsity pattern. Therefore, one can expect that the elastic net-regularized maps could possess higher robustness and stability than the L_1-norm regularized maps.

Interpretation harmonization. The purpose of interpretation harmonization is to align the saliency map generated by the neural network with some reference attention map. The reference attention map can be manually labeled by domain experts, or based on some physical principle behind the classification task. Through interpretation harmonization, the network could be enabled to analyze the phenomenon in a similar way as a human, which would help diminish the potential bias or shortcuts learned by the neural net [6, 7].

To achieve this goal, we leverage L_2-norm regularization to diminish the importance scores of unrelated features. Specifically, assume $\mathbf{A} \in \mathbb{R}^d$ is the expert's attention map with non-negative entries. A greater value implies a more important feature. We define $\tilde{\mathbf{A}} := \max(\mathbf{A}) - \mathbf{A} + \sigma$, where σ is small enough so that the ordinal relation is reverted and all elements are positive. To diminish the importance score of the background region, we use the penalty term:

$$h^\star(\mathbf{x}) = \epsilon\big\|\sqrt{\tilde{\mathbf{A}}} \odot \mathbf{x}\big\|_2^2, \epsilon > 0$$

where $\odot$ denotes Hadamard product. By adding this regularization term, the importance score of unrelated features is pushed to 0 while important features remain unaffected.

Proposition 4 *The Fenchel conjugate of the above weighted L_2-norm square regularization is:*

$$h(\mathbf{z}) = \frac{1}{4\epsilon}\Big\|\frac{1}{\sqrt{\tilde{\mathbf{A}}}} \odot \mathbf{z}\Big\|_2^2.$$

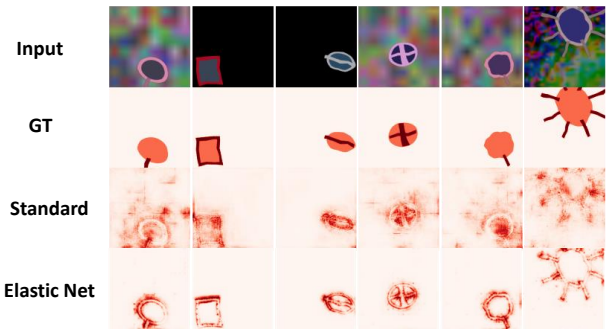

Figure 2. Qualitative results based on the synthesized dataset.

The derivation requires the weight term $\tilde{\mathbf{A}}$ to be positive. However, a positive weight only leads to the diminishing of the importance score of unrelated features. It does not explicitly promote larger scores for important features. Empirically, we have found that using δ^* in the opposite direction to the gradients for important pixels helps increase the attribution scores for important features and results in better alignment. Therefore, in our numerical experiment, we set $\tilde{\mathbf{A}} = \frac{1}{2}\max(\mathbf{A}) - \mathbf{A}$ and $\delta^* = 2\epsilon\tilde{\mathbf{A}} \odot \nabla_{\mathbf{x}}\mathcal{L}(f_\theta(\mathbf{x}), y)$.

5. Experiments

We conducted comprehensive numerical experiments to evaluate the performance of our proposed adversarial training-based methods. Firstly, we evaluate saliency maps' accuracy on a synthesized dataset with ground truth interpretation. Next, we assessed the visual quality, sparsity, interpretability, robustness, and stability of the saliency maps on Imagenette. Finally, we demonstrated the effectiveness of our interpretation harmonization strategy on the CUB-GHA dataset. Our quantitative results include adversarial training with one normalized step and multi-step gradient ascent. The main text focuses on the results of one normalized step, while detailed training settings and complete numerical results can be found in the Appendix.

5.1. Results on synthesized dataset

To illustrate how adversarial training can improve the model's interpretability, we first conducted experiments on a synthesized dataset with ground-truth for the saliency maps proposed by [35]. There are three regions on the ground-truth of the synthesized dataset (Fig. 2): background without any classification information (white region); localization information describes the location of an object (light red region); and distinguishing features crucial for classification (dark red area). We trained ResNet-34 [12] with standard training and regularized adversarial training.

As the ground-truth was a ternary class, we reported the five-band-score reflects pixel accuracy, recall, precision, and the false positive rate (FPR) based on ternary classifi-

Table 2. Comparison of five-band-scores and binary scores of standard training and adversarial training.

Methods	Five-band-scores				Binary scores			
	Pixel_acc (↑)	Recall (↑)	Precision (↑)	FPR (↓)	Pixel_acc (↑)	Recall (↑)	Precision (↑)	FPR (↓)
standard	89.81	9.86	32.58	1.21	88.66	44.47	16.22	10.22
L_1-norm	89.82	6.41	29.20	**0.76**	94.60	65.06	30.84	4.70
group-norm	89.84	7.54	27.35	0.91	94.27	63.78	29.77	4.99
elastic net	89.91	8.35	26.83	0.92	**94.63**	65.65	**31.29**	**4.66**
harmonization	**90.14**	**11.37**	**43.31**	1.14	93.82	**67.05**	28.15	5.54

Figure 3. Images optimized for maximizing class-logit activations.

cation. The results are listed in Table 2, where the method's name corresponds to the penalty term $h^\star$ in Eq. 3. The harmonization training achieved the best scores in terms of most of the metrics, showing the effectiveness of the proposed harmonization method. We also noticed that adversarial training-based methods consistently achieved higher accuracy and lower FPR compared with stranded training but led to lower recall and precision, because the adversarial training raised the sparsity of gradient maps. The importance scores of the localization area, which was less important compared with distinguishing features, were also diminished. Therefore, following [11], we also converted the ternary classification task to a binary classification task by considering the background and localization information as one class. The results suggested that adversarial training could significantly improve the accuracy of saliency maps in detecting distinguishing features.

We visualized the saliency maps in Fig. 2 for qualitative evaluation (we only show results with elastic net regularization here and leave the rest in the Appendix). We discovered that the saliency map of standard training was activated more clearly in the region of the localization information than it was in the region of the distinguishing feature. In contrast, the saliency map with adversarial training was activated more significantly for the distinguishing features, especially on the boundaries. Moreover, the saliency maps looked more sparse and less noisy.

5.2. Results on ImageNette

We performed numerical experiments on ImageNette, a ten-class subset of ImageNet [4], to examine the properties of the proposed adversarial training. Specifically, we studied the sparsity, interpretability, adversarial robustness, and stability of the saliency maps generated for EfficientNet [34] classifiers with the proposed adversarial training methods.

Sparsity. We trained the neural network models with different regularization coefficients and used the Gini Index [13] as a measure of the sparsity of the saliency map. The results are shown in Table 3. All the proposed adversarial training-based methods led to more sparse saliency maps than standard training, with zero or only minor drops in clean accuracy. The L_1-norm-based method attained the highest sparsity. Also, the sparsity increased with the regularization coefficient ϵ. The gradient maps visualized in Fig. 4 suggest that standard training could lead to miscellaneous points in the background. On the other hand, the gradient maps with our proposed adversarial training methods seem to have higher visual quality.

Interpretability. To show that the proposed training methods can improve the interpretability of the model, we calculated the DiffROAR score [26] for the methods. DiffROAR measures the difference in the predictive power of the dataset, where the top and bottom $k\%$ of the pixels are removed according to the importance score. A higher DiffROAR score shows the model captures the task-related features from the inputs. To this end, we measured the DiffROAR score for each method with k from 10% to 90% in increments of 10%. Each measurement was repeated with three different initializations, and the final DiffROAR score was the average of the 27 measurements. The results are shown in Table 3. Compared with standard training, the proposed methods can improve the model's interpretability, especially under the elastic net regularization, which achieved the highest DiffROAR score.

We also visualized the learned feature for each model, using the optimization-based feature visualization technique of *deep-dream* [20]. The results (Fig. 3) also show that class-specific features identified with the proposed norm-regularized adversarial training methods look more meaningful to human perception.

Robustness. To assess the robustness of interpretation maps under the standard and proposed training methods, we adopted an L_2-norm bounded interpretation attack method proposed by [18]. We gradually increased the strength of the attack and measured the robustness through similarity measures of the saliency maps before and after the attack. We adopted two metrics as robustness measures: 1) top-k intersection ratio [8], which is the ratio of pix-

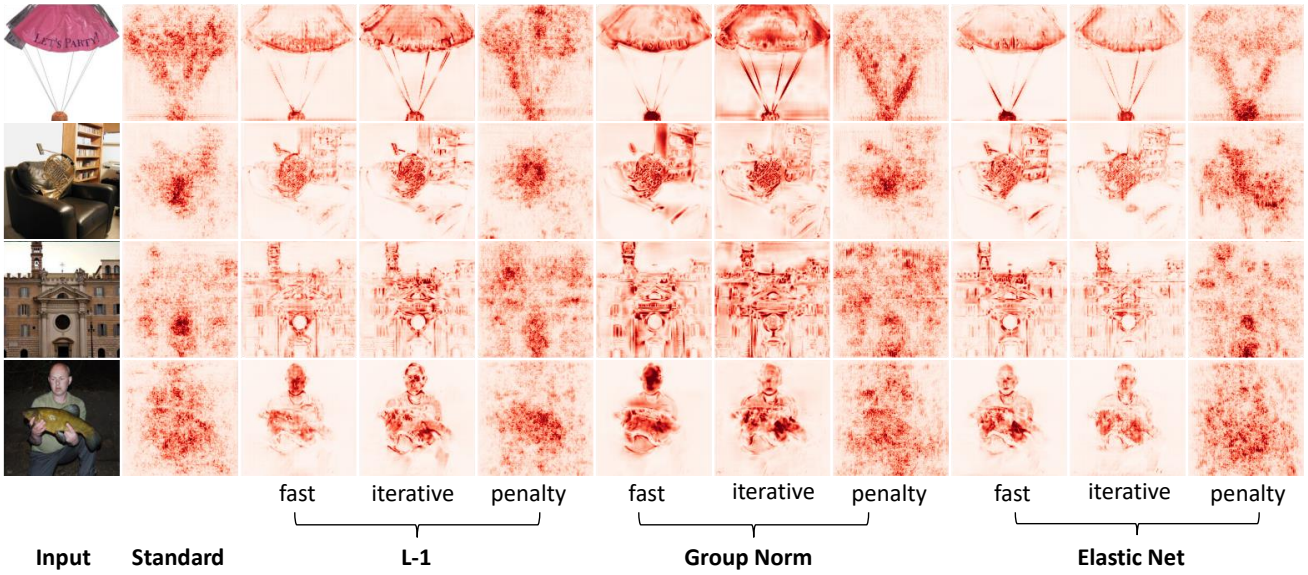

Figure 4. Qualitative comparison of saliency maps generated by networks with different adversarial training protocols (fast, iterative) and standard training with additional regularization on input gradients (penalty).

Table 3. Quantitative evaluation on ImageNette.

Methods	DiffROAR (%) (↑)	Gini (%) (↑)	Acc. (%) (↑)
standard	1.81	46.53	89.04
L_1-norm (ϵ=0.01)	2.95	56.73	**90.46**
L_1-norm (ϵ=0.05)	2.12	62.04	84.13
L_1-norm (ϵ=0.10)	1.97	**65.53**	78.85
group-norm (ϵ=0.10)	2.57	50.07	86.83
group-norm (ϵ=0.50)	1.32	49.15	87.57
group-norm (ϵ=1.00)	0.40	52.70	83.80
elastic net (ϵ_1=0.01, ϵ_2=0.01)	2.60	57.10	87.72
elastic net (ϵ_1=0.01, ϵ_2=0.05)	**3.24**	57.35	87.67
elastic net (ϵ_1=0.01, ϵ_2=0.10)	3.14	54.44	87.19
elastic net (ϵ_1=0.05, ϵ_2=0.01)	2.31	60.05	84.59
elastic net (ϵ_1=0.05, ϵ_2=0.05)	2.11	61.84	84.69
elastic net (ϵ_1=0.05, ϵ_2=0.10)	2.17	60.25	84.54

els that remain salient after the interpretation attack, and 2) the structural similarity index measure (SSIM) [38]. We compared the proposed training methods with the baseline standard training and three post-processing-based baselines, SmoothGrad, Sparsified-SmoothGrad (Sparsified-S.), and MoreauGrad (Fig. 5). Standard training seemed vulnerable to minor attacks, even when combined with the Smooth-based post-processing. In contrast, the neural nets trained by the proposed methods displayed significantly higher robustness to the attacks. We also observed that the elastic net and group-norm-based training methods performed slightly more robustly than the L_1-norm-based method.

Stability. We evaluated the algorithmic stability as the magnitude of perturbation in the saliency maps under classifiers trained with different training data or random initialization. As pointed out by [39], standard training with different initializations can result in large discrepancies in saliency maps. We examined if the proposed training methods can alleviate this issue. To this end, we switched 10% of the training data with the test data and initialized two networks with different random seeds. Then we trained these two networks with the same training process. We used SSIM and top-k overlap to measure the similarities of the generated saliency maps, where the top-k overlap was defined as the Dice score of the top-k masks. As shown in Fig. 5, adversarial training could significantly improve stability. It even brings more improvement than post-processing methods. Note for Sparsified-SmoothGrad, the high SSIM is due to the sparsity of the saliency maps. Through studying top-k overlap, the activated regions were actually different. Among the three variants of adversarial training, the elastic net could achieve the best stability, as the incorporation of L_2^2 term could increase the smoothness of the objective function, making the optimization process less sensitive to the perturbation of the training data.

Comparison of different training protocols. We also conducted experiments to see how different training protocols affect the numerical results. Specifically, we compared one-step optimization (fast), multiple-step optimization (iterative), and free adversarial training (free, in Appendix) [25]. As a baseline, we also studied standard training with norm regularization directly applied to the objective function (penalty). We visualized a few saliency maps in Fig. 4 for comparison. Standard training with regularization failed to generate structured saliency maps, showing the challenge of network optimization with regularization in the objective function, which further certified the superiority of adversarial training. Visually, fast, and iterative

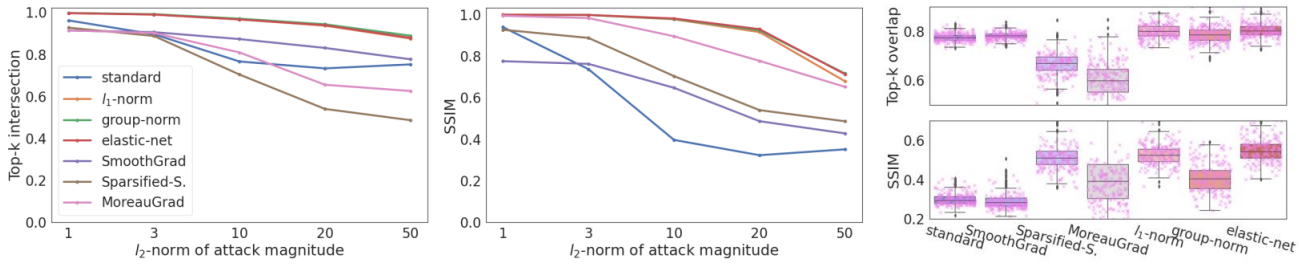

Figure 5. Quantitative robustness and stability comparison. **Left/Middle**: Comparison of SSIM/top-k intersection of saliency maps before and after the attack. **Right**: Comparison of SSIM/top-k overlap of saliency maps generated by networks with different stochastic training.

Table 4. Top-k overlap (%) between gaze map and saliency map on CUB-GHA dataset.

ϵ	0.0	0.1	0.5	1.0	5.0
top 5% overlap	33.42	35.07	38.74	42.63	50.62
top 10% overlap	41.81	42.99	45.60	47 06	52.60
accuracy (%)	67.58	67.50	66.56	66.65	64.99

training achieved comparable quality. However, for L_1-norm, interactive training would lose sparsity compared to fast training, but for the elastic net, the trend was the opposite. The Fenchel conjugate of elastic net regularization is differentiable everywhere due to the square term, making its optimization easier compared with L_1-norm.

5.3. Harmonization with gaze maps

We conducted experiments on the CUB-GHA [21] to verify the effectiveness of our interpretation harmonization strategy on real-world data. CUB-GHA is an extension of CUB [36], which includes gaze maps collected from domain experts for bird category classification. The gaze map is a type of human attention map reflecting how human experts would conduct this task. We aim to harmonize the saliency maps with human attention so that the network could make the decision more similar to humans.

To do this, we normalized the gaze map and used the attention score as the weights for the l_2-norm perturbation during adversarial training. The results are shown in Fig. 6. The original saliency maps highlighted the whole body of the bird, while the gaze maps were more focused on a specific part (e.g. head) of the bird. After harmonization, the saliency maps showed better alignment with the gaze maps. We also calculated the top-k overlap between saliency maps and gaze maps (Table 4). The results showed an increasing trend of the overlap as we raised the coefficient ϵ, while the trade-off of accuracy was in a reasonable range.

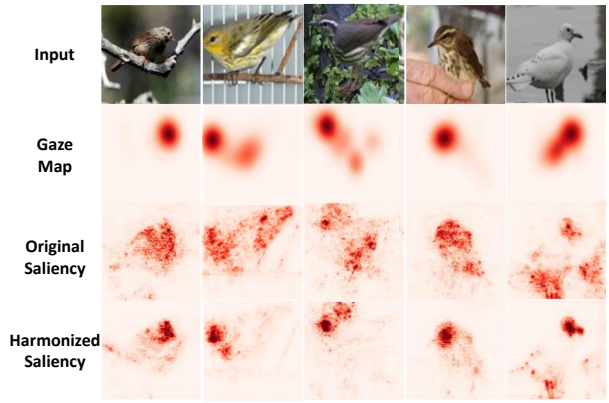

Figure 6. Qualitative results based on CUB-GHA dataset.

6. Conclusion

In this paper, we introduced a duality framework to analyze the impact of norm-regularized adversarial training on the gradient-based saliency maps of a trained neural net classifier. Leveraging this duality framework, we proposed several variants of norm-regularized adversarial training, designed to promote certain structures in the gradient-based interpretation maps, including sparsity, group sparsity, and consistency with human attention. We provided experimental results on several benchmark datasets to validate the effectiveness of our proposed methods, which demonstrated that properly designed adversarial training methods can enhance model interpretability as well as the stability and robustness of the gradient-based maps. An interesting future direction is to develop a similar adversarial training-based regularization framework for structured interpretation maps according to integrated gradients, DeepLIFT, and Grad-Cam. Another relevant future direction is to explore other forms of penalty terms in order to expand the methodology to domains beyond computer vision.

Acknowledgments. The work of Farzan Farnia is partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China, Project 14209920, and is partially supported by a CUHK Direct Research Grant with CUHK Project No. 4055164.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018. 2, 4

[2] Nishanth Arun, Nathan Gaw, Praveer Singh, Ken Chang, Mehak Aggarwal, Bryan Chen, Katharina Hoebel, Sharut Gupta, Jay Patel, Mishka Gidwani, et al. Assessing the trustworthiness of saliency maps for localizing abnormalities in medical imaging. *Radiology: Artificial Intelligence*, 3(6): e200267, 2021. 2

[3] Prasad Chalasani, Jiefeng Chen, Amrita Roy Chowdhury, Xi Wu, and Somesh Jha. Concise explanations of neural networks using adversarial training. In *International Conference on Machine Learning*, pages 1383–1391. PMLR, 2020. 3

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[5] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018. 3

[6] Thomas Fel, Ivan F Rodriguez Rodriguez, Drew Linsley, and Thomas Serre. Harmonizing the object recognition strategies of deep neural networks with humans. *Advances in Neural Information Processing Systems*, 35:9432–9446, 2022. 5

[7] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. 5

[8] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3681–3688, 2019. 6

[9] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 2

[10] Yanming Guo, Yu Liu, Theodoros Georgiou, and Michael S Lew. A review of semantic segmentation using deep neural networks. *International journal of multimedia information retrieval*, 7:87–93, 2018. 1

[11] Hyeongrok Han, Siwon Kim, Hyun-Soo Choi, and Sungroh Yoon. On the impact of knowledge distillation for model interpretability. *arXiv preprint arXiv:2305.15734*, 2023. 6

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5, 2

[13] Niall Hurley and Scott Rickard. Comparing measures of sparsity. *IEEE Transactions on Information Theory*, 55(10): 4723–4741, 2009. 6

[14] Beomsu Kim, Junghoon Seo, Seunghyeon Jeon, Jamyoung Koo, Jeongyeol Choe, and Taegyun Jeon. Why are saliency maps noisy? cause of and solution to noisy saliency maps. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 4149–4157. IEEE, 2019. 3

[15] Beomsu Kim, Junghoon Seo, and Taegyun Jeon. Bridging adversarial robustness and gradient interpretability. *arXiv preprint arXiv:1903.11626*, 2019. 3

[16] Jinkyu Kim and John Canny. Interpretable learning for self-driving cars by visualizing causal attention. In *Proceedings of the IEEE international conference on computer vision*, pages 2942–2950, 2017. 1

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 1

[18] Alexander Levine, Sahil Singla, and Soheil Feizi. Certifiably robust interpretation in deep learning. *arXiv preprint arXiv:1905.12105*, 2019. 2, 6

[19] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 2

[20] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2(11):e7, 2017. 6

[21] Yao Rong, Wenjia Xu, Zeynep Akata, and Enkelejda Kasneci. Human attention in fine-grained classification. *arXiv preprint arXiv:2111.01628*, 2021. 8

[22] Andrew Ross and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 3

[23] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673, 2016. 4

[24] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 2

[25] Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! *Advances in Neural Information Processing Systems*, 32, 2019. 7, 3

[26] Harshay Shah, Prateek Jain, and Praneeth Netrapalli. Do input gradients highlight discriminative features? *Advances in Neural Information Processing Systems*, 34:2046–2059, 2021. 2, 3, 6

[27] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19:221–248, 2017. 1

[28] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR, 2017. 2

[29] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 1, 2

[30] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017. 2, 3

[31] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014. 2

[32] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017. 1, 2

[33] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 2

[34] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 6, 2

[35] Erico Tjoa and Guan Cuntai. Quantifying explainability of saliency methods in deep neural networks with a synthetic dataset. *IEEE Transactions on Artificial Intelligence*, 2022. 5

[36] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 8

[37] Mei Wang and Weihong Deng. Deep face recognition: A survey. *Neurocomputing*, 429:215–244, 2021. 1

[38] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 7

[39] Ann-Christin Woerl, Jan Disselhoff, and Michael Wand. Initialization noise in image gradients and saliency maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1766–1775, 2023. 7

[40] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I Inouye, and Pradeep K Ravikumar. On the (in) fidelity and sensitivity of explanations. *Advances in Neural Information Processing Systems*, 32, 2019. 4

[41] Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(1):49–67, 2006. 2

[42] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer, 2014. 2

[43] Jingwei Zhang and Farzan Farnia. Moreaugrad: Sparse and robust interpretation of neural networks via moreau envelope. *arXiv preprint arXiv:2302.05294*, 2023. 2, 4

[44] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems*, 30 (11):3212–3232, 2019. 1

[45] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005. 2