MoreauGrad: Sparse and Robust Interpretation of Neural Networks via Moreau Envelope

Jingwei Zhang
The Chinese University of Hong Kong
jwzhang22@cse.cuhk.edu.hk

Farzan Farnia
The Chinese University of Hong Kong
farnia@cse.cuhk.edu.hk

Abstract

Explaining the predictions of deep neural nets has been a topic of great interest in the computer vision literature. While several gradient-based interpretation schemes have been proposed to reveal the influential variables in a neural net's prediction, standard gradient-based interpretation frameworks have been commonly observed to lack robustness to input perturbations and flexibility for incorporating prior knowledge of sparsity and group-sparsity structures. In this work, we propose MoreauGrad as an interpretation scheme based on the classifier neural net's Moreau envelope. We demonstrate that MoreauGrad results in a smooth and robust interpretation of a multi-layer neural network and can be efficiently computed through first-order optimization methods. Furthermore, we show that MoreauGrad can be naturally combined with L_1-norm regularization techniques to output a sparse or group-sparse explanation which are prior conditions applicable to a wide range of deep learning applications. We empirically evaluate the proposed MoreauGrad scheme on standard computer vision datasets, showing the qualitative and quantitative success of the MoreauGrad approach in comparison to standard gradient-based interpretation methods [1].

1. Introduction

Deep neural networks (DNNs) have achieved state-of-the-art performance in many computer vision problems including image classification [11], object detection [29], and medical image analysis [19]. While they manage to attain super-human scores on standard image and speech recognition tasks, a reliable application of deep learning models to real-world problems requires an interpretation of their predictions to help domain experts understand and investigate the basis of their predictions. Over the past few years, developing and analyzing interpretation schemes that reveal the influential features in a neural network's prediction have attracted great interest in the computer vision community.

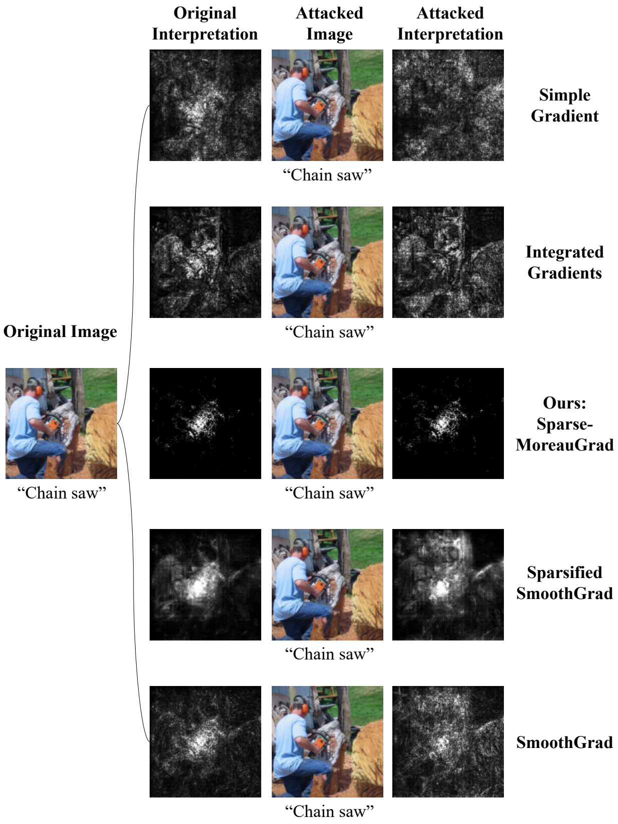

Figure 1. Interpretation of Sparse MoreauGrad (ours) vs. standard gradient-based baselines on an ImageNet sample before and after adding a norm-bounded interpretation adversarial attack.

A standard approach for interpreting neural nets' predictions is to analyze the gradient of their prediction score function at or around an input data point. Such gradient-based interpretation mechanisms result in a feature saliency map revealing the influential variables that locally affect the neural net's assigned prediction score. Three well-known examples of gradient-based interpretation schemes are the simple gradient [21], integrated gradients [24], and

[1] The paper's code is available at https://github.com/buyeah1109/MoreauGrad.

DeepLIFT [20] methods. While the mentioned methods have found many applications in explaining neural nets' predictions, they have been observed to lack robustness to input perturbations and to output a dense noisy saliency map in their application to computer vision datasets [6, 8]. Consequently, these gradient-based explanations can be considerably altered by minor random or adversarial input noise.

A widely-used approach to improve the robustness and sharpness of gradient-based interpretations is SmoothGrad [22] which applies Gaussian smoothing to the mentioned gradient-based interpretation methods. As shown by [22], SmoothGrad can significantly boost the visual quality of a neural net's gradient-based saliency map. On the other hand, SmoothGrad typically leads to a dense interpretation vector and remains inflexible to incorporate prior knowledge of sparsity and group-sparsity structures. Since a sparse saliency map is an applicable assumption to several image classification problems where a relatively small group of input variables can completely determine the image label, a counterpart of SmoothGrad which can simultaneously achieve sparse and robust interpretation will be of significant use in computer vision problems.

In this paper, we propose a novel approach, which we call *MoreauGrad*, to achieve a provably smooth gradient-based interpretation with potential sparsity or group-sparsity properties. The proposed MoreauGrad outputs the gradient of a classifier's Moreau envelope which is a useful optimization tool for enforcing smoothness in a target function. We leverage convex analysis to show that MoreauGrad behaves smoothly around an input sample and therefore provides an alternative optimization-based approach to SmoothGrad for achieving a smoothly-changing saliency map. As a result, we demonstrate that similar to SmoothGrad, MoreauGrad offers robustness to input perturbations, since a norm-bounded perturbation will only lead to a bounded change to the MoreauGrad interpretation.

Next, we show that MoreauGrad can be flexibly combined with L_1-norm-based regularization penalties to output sparse and group-sparse interpretations. Our proposed combinations, Sparse MoreauGrad and Group-Sparse MoreauGrad, take advantage of elastic-net [31] and group-norm [16] penalty terms to enforce sparse and group-sparse saliency maps, respectively. We show that these extensions of MoreauGrad preserve the smoothness and robustness properties of the original MoreauGrad scheme. Therefore, our discussion demonstrates the adaptable nature of MoreauGrad for incorporating prior knowledge of sparsity structures in the output interpretation.

Finally, we present the empirical results of our numerical experiments applying MoreauGrad to standard image recognition datasets and neural net architectures. We compare the numerical performance of MoreauGrad with standard gradient-based interpretation baselines. Our numerical results indicate the satisfactory performance of vanilla and L_1-norm-based MoreauGrad in terms of visual quality and robustness. Figure 1 shows the robustness and sparsity of the Sparse MoreauGrad interpretation applied to an ImageNet sample in comparison to standard gradient-based saliency maps. As this and our other empirical findings suggest, MoreauGrad can outperform standard baselines in terms of the sparsity and robustness properties of the output interpretation. In the following, we summarize the main contributions of this paper:

- Proposing MoreauGrad as an interpretation scheme based on a classifier function's Moreau envelope
- Analyzing the smoothness and robustness properties of MoreauGrad by leveraging convex analysis
- Introducing L_1-regularized Sparse MoreauGrad to obtain an interpretation satisfying prior sparsity conditions
- Providing numerical results supporting MoreauGrad over standard image recognition datasets

2. Related Work

Gradient-based Interpretation. A large body of related works develop gradient-based interpretation methods. Simonyan et al. [21] propose to calculate the gradient of a classifier's output with respect to an input image. The simple gradient approach in [21] has been improved by several related works. Notably, the method of Integrated Gradients [24] is capable of keeping highly relevant pixels in the saliency map by aggregating gradients of image samples. SmoothGrad [22] removes noise in saliency maps by adding Gaussian-random noise to the input image. The CAM method [30] analyzes the information from global average pooling layer for localization, and Grad-CAM++ [1] improves over Grad-CAM [18] and generates coarse heatmaps with improved multi-object localization. The NormGrad [17] focuses on the weight-based gradient to analyze the contribution of each image region. DeepLIFT [20] uses difference from reference to propagate an attribution signal. However, the mentioned gradient-based methods do not obtain a sparse interpretation, and their proper combination with L_1-regularization to promote sparsity remains highly non-trivial and challenging. On the other hand, our proposed MoreauGrad can be smoothly equipped with L_1-regularization to output sparse interpretations and can further capture group-sparsity structures.

Mask-based Interpretation. Mask-based interpretation methods rely on adversarial perturbations to interpret neural nets. Applying a mask which perturbs the neural net input, the importance of input pixels is measured by a masked-based method. This approach to explaining neural nets has been successfully applied in References [26, 5, 14, 2] and has been shown to benefit from dynamic perturbations [9]. More specifically, Dabkowski and Gal [2] introduce a real-time mask-based detection method; Fong and Vedaldi [5] develop a model-agnostic approach with interpretable perturbations; Wagner et al. [26] propose a method that could generate fine-grained visual interpretations. Moreover, Lim et al. [14] leverage local smoothness to enhance their robustness towards samples attacked by PGD [15]. However, [5] and [2] show that perturbation-based interpretation methods are still vulnerable to adversarial perturbations.

We note that the discussed methods depend on optimizing perturbation masks for interpretations, and due to the non-convex nature of neural net loss functions, their interpretation remains sensitive to input perturbations. In contrast, our proposed MoreauGrad can provably smooth the neural net score function, and can adapt to non-convex functions using norm regularization. Hence, MoreauGrad can improve both the sparsity and robustness of the interpretation.

Robust Interpretation. The robustness of interpretation methods has been a subject of great interest in the literature. Ghorbani et al. [6] introduce a gradient-based adversarial attack method to alter the neural nets' interpretation. Dombrowski et al. [4] demonstrate that interpretations could be manipulated, and they suggest improving the robustness via smoothing the neural net classifier. Heo et al. [8] propose a manipulation method that is capable of generalizing across datasets. Subramanya et al. [23] create adversarial patches fooling both the classifier and the interpretation.

To improve the robustness, Sparsified-SmoothGrad [13] combines a sparsification technique with Gaussian smoothing to achieve certifiable robustness. The related works [26, 28, 5, 14, 2] discuss the application of adversarial defense methods against classification-based attacks to interpret the prediction of neural net classifiers. We note that these papers' main focus is not on defense schemes against interpretation-based attacks. Specifically, [26] filter gradients internally during backpropogation, and [14] leverage local smoothness to integrate more samples. Unlike the mentioned papers, our work proposes a model-agnostic optimization-based method which is capable of generating simultaneously sparse and robust interpretations.

3. Preliminaries

In this section, we review three standard interpretation methods as well as the notation and definitions in the paper.

3.1. Notation and Definitions

In the paper, we use notation $\mathbf{X} \in \mathbb{R}^d$ to denote the feature vector and $Y \in \{1,\ldots,k\}$ to denote the label of a sample. In addition, $f_{\mathbf{w}} : \mathbb{R}^d \rightarrow \mathbb{R}^k$ denotes a neural net classifier with its weights contained in vector $\mathbf{w} \in \mathcal{W}$ where $\mathcal{W}$ is the feasible set of the neural net's weights. Here $f_{\mathbf{w}}$ maps the d-dimensional input $\mathbf{x}$ to a k-dimensional prediction vector containing the likelihood of each of the k classes in the classification problem. For every class $c \in \{1,\ldots,k\}$, we use the notation $f_{\mathbf{w},c} : \mathbb{R}^d \rightarrow \mathbb{R}$ to denote the c-th entry of $f_{\mathbf{w}}$'s output which corresponds to class c.

We use $\|\mathbf{x}\|_p$ to denote the ℓ_p-norm of input vector $\mathbf{x}$. Furthermore, we use notation $\|\mathbf{x}\|_{p,q}$ to denote the $\ell_{p,q}$-group-norm of $\mathbf{x}$ defined in the following equation for given variable subsets $S_1,\ldots,S_t \subseteq \{1,\ldots,d\}$:

$$\|\mathbf{x}\|_{p,q} = \Big\| \big[\|\mathbf{x}_{S_1}\|_p,\ldots,\|\mathbf{x}_{S_t}\|_p\big] \Big\|_q \tag{1}$$

In other words, $\|\mathbf{x}\|_{p,q}$ is the ℓ_q-norm of a vector containing the ℓ_p-norms of the subvectors of $\mathbf{x}$ characterized by index subsets $S_1,\ldots,S_t$.

3.2. Gradient-based Saliency Maps

In our theoretical and numerical analysis, we consider the following widely-used gradient-based interpretation baselines which apply to a classifier neural net $f_{\mathbf{w}}$ and predicted class c for input $\mathbf{x}$:

1. **Simple Gradient**: The simple gradient interpretation returns the saliency map of a neural net score function's gradient with respect to input $\mathbf{x}$:

$$\mathrm{SG}\big(f_{\mathbf{w},c},\mathbf{x}\big) := \nabla_{\mathbf{x}} f_{\mathbf{w},c}(\mathbf{x}). \tag{2}$$

   In the applications of the simple gradient approach, c is commonly chosen as the neural net's predicted label with the maximum prediction score.

2. **Integrated Gradients:** The integrated gradients approach approximates the integral of the neural net's gradient function between a reference point $\mathbf{x}^0$ and the input $\mathbf{x}$. Using m intermediate points on the line segment connecting $\mathbf{x}^0$ and $\mathbf{x}$, the integrated gradient output will be

$$\mathrm{IG}\big(f_{\mathbf{w},c},\mathbf{x}\big) := \frac{\Delta\mathbf{x}}{m}\sum_{i=1}^{m} \nabla_{\mathbf{x}} f_{\mathbf{w},c}\big(\mathbf{x}^0 + \frac{i}{m}\Delta\mathbf{x}\big). \tag{3}$$

   In the above $\Delta\mathbf{x} := \mathbf{x} - \mathbf{x}^0$ denotes the difference between the target and reference points $\mathbf{x}, \mathbf{x}^0$.

3. **SmoothGrad:** SmoothGrad considers the averaged simple gradient score over an additive random perturbation Z drawn according to an isotropic Gaussian distribution $Z \sim \mathcal{N}(\mathbf{0},\sigma^2 I_d)$. In practice, the SmoothGrad interpretation is estimated over a number t of independently drawn noise vectors $\mathbf{z}_1,\ldots,\mathbf{z}_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0},\sigma^2 I_d)$ according to the zero-mean Gaussian distribution:

$$\begin{aligned}\mathrm{SmoothGrad}\big(f_{\mathbf{w},c},\mathbf{x}\big) &:= \mathbb{E}\big[\nabla_{\mathbf{x}} f_{\mathbf{w},c}(\mathbf{x}+Z)\big] \\ &\approx \frac{1}{t}\sum_{i=1}^{t} \nabla_{\mathbf{x}} f_{\mathbf{w},c}(\mathbf{x}+\mathbf{z}_i).\end{aligned} \tag{4}$$

4. MoreauGrad: An Optimization-based Interpretation Framework

As discussed earlier, smooth classifier functions with a Lipschitz gradient help to obtain a robust explanation of neural nets. Here, we propose an optimization-based smoothing approach based on Moreau-Yosida regularization. To introduce this optimization-based approach, we first define a function's Moreau envelope.

Definition 1. *Given regularization parameter $\rho > 0$, we define the Moreau envelope of a function $g : \mathbb{R}^d \rightarrow \mathbb{R}$ as:*

$$g^{\rho}(\mathbf{x}) := \min_{\widetilde{\mathbf{x}} \in \mathbb{R}^d} \, g\bigl(\widetilde{\mathbf{x}}\bigr) + \frac{1}{2\rho}\bigl\|\widetilde{\mathbf{x}} - \mathbf{x}\bigr\|_2^2. \tag{5}$$

In the above definition, $\rho > 0$ represents the Moreau-Yosida regularization coefficient. Applying the Moreau envelope, we propose the MoreauGrad interpretation as the gradient of the classifier's Moreau envelope at an input $\mathbf{x}$.

Definition 2. *Given regularization parameter $\rho > 0$, we define the MoreauGrad interpretation $MG_\rho : \mathbb{R}^d \rightarrow \mathbb{R}^d$ of a neural net $f_{\mathbf{w}}$ predicting class c for input $\mathbf{x}$ as*

$$MG_\rho(f_{\mathbf{w},c}, \mathbf{x}) := \nabla f^{\rho}_{\mathbf{w},c}(\mathbf{x}).$$

To compute and analyze the MoreauGrad explanation, we first discuss the optimization-based smoothing enforced by the Moreau envelope. Note that the Moreau envelope is known as an optimization tool to turn non-smooth convex functions (e.g. ℓ_1-norm) into smooth functions. Here, we discuss an extension of this result to weakly-convex functions which also apply to non-convex functions.

Definition 3. *A function $g : \mathbb{R}^d \rightarrow \mathbb{R}$ is called λ-weakly convex if $\Phi(\mathbf{x}) := g(\mathbf{x}) + \frac{\lambda}{2}\|\mathbf{x}\|_2^2$ is a convex function, i.e. for every $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^d$ and $0 \le \alpha \le 1$ we have:*

$$\begin{aligned} g\bigl(\alpha\mathbf{x}_1 + (1-\alpha)\mathbf{x}_2\bigr) \; &\le \alpha g(\mathbf{x}_1) + (1-\alpha) g(\mathbf{x}_2) \\ &\quad + \frac{\lambda\alpha(1-\alpha)}{2}\bigl\|\mathbf{x}_1 - \mathbf{x}_2\bigr\|_2^2. \end{aligned}$$

Theorem 1. *Suppose that $g : \mathbb{R}^d \rightarrow \mathbb{R}$ is a λ-weakly convex function. Assuming that $0 < \rho < \frac{1}{\lambda}$, the followings hold for the optimization problem of the Moreau envelope g^ρ and the optimal solution $\widetilde{x}^*_\rho(\mathbf{x})$ solving the optimization problem:*

1. *For every $\mathbf{x}$, the gradient of g^ρ and Clarke subdifferential of g (denoted by $\partial_0 g$) are related as:*

   $$\nabla g^{\rho}(\mathbf{x}) \in \partial_0 g\bigl(\widetilde{x}^*_\rho(\mathbf{x})\bigr),$$

   which implies that if g is differentiable at $\widetilde{x}^*_\rho(\mathbf{x})$ we have $\nabla g^\rho(\mathbf{x}) = \nabla g(\widetilde{x}^*_\rho(\mathbf{x}))$.

2. *The difference $\widetilde{x}^*_\rho(\mathbf{x}) - \mathbf{x}$ is aligned with g^ρ's gradient:*

   $$\nabla g^{\rho}(\mathbf{x}) = \frac{-1}{\rho}\bigl(\,\widetilde{x}^*_\rho(\mathbf{x}) - \mathbf{x}\,\bigr).$$

3. *g^ρ will be $\max\{\frac{1}{\rho}, \frac{\lambda}{1-\rho\lambda}\}$-smooth, i.e. for every $\mathbf{x}_1, \mathbf{x}_2$:*

   $$\bigl\|\nabla g^{\rho}(\mathbf{x}_1) - \nabla g^{\rho}(\mathbf{x}_2)\bigr\|_2 \le \frac{1}{\min\{\rho, \frac{1}{\lambda} - \rho\}}\bigl\|\mathbf{x}_1 - \mathbf{x}_2\bigr\|_2.$$

Proof. This result is well-known for convex functions. In the Appendix, we provide a new proof for the result. □

Corollary 1. *Assume that the prediction score function $f_{\mathbf{w},c} : \mathbb{R}^d \rightarrow \mathbb{R}$ is λ-weakly convex. Then, the MoreauGrad interpretation MG_ρ will remain robust under an ϵ-ℓ_2-norm bounded perturbation $\|\boldsymbol{\delta}\|_2 \le \epsilon$ as*

$$\bigl\|MG_\rho(\mathbf{x} + \boldsymbol{\delta}) - MG_\rho(\mathbf{x})\bigr\|_2 \le \frac{\epsilon}{\min\{\rho, \frac{1}{\lambda} - \rho\}}.$$

The above results imply that by choosing a small enough coefficient ρ the Moreau envelope will be a differentiable smooth function. Moreover, the computation of the Moreau envelope will reduce to a convex optimization task that can be solved by standard or accelerated gradient descent with global convergence guarantees. Therefore, one can efficiently compute the MoreauGrad interpretation by solving the optimization problem via the gradient descent algorithm. Algorithm 1 applies gradient descent to compute the solution to the Moreau envelope optimization which according to Theorem 1 yields the MoreauGrad explanation.

As discussed above, MoreauGrad will be provably robust as long as the regularization coefficient will dominate the weakly-convexity degree of the prediction score. In the following proposition, we show this condition can be enforced by applying either Gaussian smoothing.

Proposition 1. *Suppose that $f_{\mathbf{w},c}$ is L-Lipschitz, that is for every $\mathbf{x}_1, \mathbf{x}_2$ $|f_{\mathbf{w},c}(\mathbf{x}_1) - f_{\mathbf{w},c}(\mathbf{x}_2)| \le L\|\mathbf{x}_2 - \mathbf{x}_1\|_2$, but could be potentially non-differentiable and non-smooth. Then, $h_{\mathbf{w},c}(\mathbf{x}) := \mathbb{E}[f_{\mathbf{w},c}(\mathbf{x} + \mathbf{Z})]$ where $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_{d\times d})$ will be $\frac{L}{\sigma}$-weakly convex.*

Proof. We postpone the proof to the Appendix. □

The above proposition suggests the regularized MoreauGrad which regularizes the neural net function to satisfy the weakly-convex condition through Gaussian smoothing.

5. Sparse and Group-Sparse MoreauGrad

To further extend the MoreauGrad approach to output sparsely-structured feature saliency maps, we further include an L_1-norm-based penalty term in the Moreau-Yosida regularization and define the following L_1-norm-based sparse and group-sparse Moreau envelope.

Definition 4. *For a function $g : \mathbb{R}^d \rightarrow \mathbb{R}$ and regularization coefficients $\rho, \eta > 0$, we define L_1-Moreau envelope $g^{\rho,\eta}_{L_1}$:*

$$g^{\rho,\eta}_{L_1}(\mathbf{x}) := \min_{\widetilde{\mathbf{x}} \in \mathbb{R}^d} \, g(\widetilde{\mathbf{x}}) + \frac{1}{2\rho}\bigl\|\widetilde{\mathbf{x}} - \mathbf{x}\bigr\|_2^2 + \eta\bigl\|\widetilde{\mathbf{x}} - \mathbf{x}\bigr\|_1.$$

We also define $L_{2,1}$-Moreau envelope $g^{\rho,\eta}_{L_{2,1}}$ as

$$g^{\rho,\eta}_{L_{2,1}}(\mathbf{x}) \;:=\; \min_{\widetilde{\mathbf{x}}\in\mathbb{R}^d}\; g(\widetilde{\mathbf{x}}) + \frac{1}{2\rho}\big\|\widetilde{\mathbf{x}}-\mathbf{x}\big\|_2^2 + \eta\big\|\widetilde{\mathbf{x}}-\mathbf{x}\big\|_{2,1}.$$

In the above, the group norm $\|\cdot\|_{2,1}$ is defined as $\|\mathbf{x}\|_{2,1} := \sum_{i=1}^{t}\|\mathbf{x}_{S_i}\|_2$ for given subsets $S_1,\dots,S_t \subseteq \{1,\dots,d\}$.

Definition 5. *Given regularization coefficients $\rho,\eta > 0$, we define the Sparse MoreauGrad (S-MG$_{\rho,\eta}$) and Group-Sparse MoreauGrad (GS-MG$_{\rho,\eta}$) interpretations as*

$$\begin{aligned}\textit{S-MG}_{\rho,\eta}(f_{\mathbf{w},c},\mathbf{x}) \;&:=\; \frac{1}{\rho}\big(\,\widetilde{\mathbf{x}}^*_{L_1}(\mathbf{x})-\mathbf{x}\,\big),\\ \textit{GS-MG}_{\rho,\eta}(f_{\mathbf{w},c},\mathbf{x}) \;&:=\; \frac{1}{\rho}\big(\,\widetilde{\mathbf{x}}^*_{L_{2,1}}(\mathbf{x})-\mathbf{x}\,\big),\end{aligned}$$

where $\widetilde{\mathbf{x}}^*_{L_1}(\mathbf{x})$, $\widetilde{\mathbf{x}}^*_{L_{2,1}}(\mathbf{x})$ denote the optimal solutions to the optimization tasks of $f^{\rho,\eta}_{\mathbf{w},c,L_1}(\mathbf{x})$, $f^{\rho,\eta}_{\mathbf{w},c,L_{2,1}}(\mathbf{x})$, respectively.

In the following theorem, we extend the shown results for the smoothness and robustness of vanilla Moreau envelope to our proposed L_1-norm-based extensions of the Moreau envelope. In this theorem, we use the following definitions for the soft-thresholding operators ST$_\alpha$ and GST$_\alpha$ for the sparse and group-sparse cases which are defined entrywise and group-entrywise as follows

$$\mathrm{ST}_\alpha(\mathbf{x})_i := \begin{cases} 0 & \text{if } |x_i| \le \alpha \\ x_i - \mathrm{sign}(x_i)\alpha & \text{if } |x_i| > \alpha, \end{cases}$$

$$\mathrm{GST}_\alpha(\mathbf{x})_{S_i} := \begin{cases} \mathbf{0} & \text{if } \|\mathbf{x}_{S_i}\|_2 \le \alpha \\ \big(1-\frac{\alpha}{\|\mathbf{x}_{S_i}\|_2}\big)\mathbf{x}_{S_i} & \text{if } \|\mathbf{x}_{S_i}\|_2 > \alpha. \end{cases}$$

Theorem 2. *Suppose that $g:\mathbb{R}^d\rightarrow\mathbb{R}$ is a λ-weakly convex function. Then, assuming that $0<\rho<\frac{1}{\lambda}$, Theorem 1's parts 1 and 3 will further hold for the sparse Moreau envelope $g^{\rho,\eta}_{L_1}$ and group-sparse Moreau envelope $g^{\rho,\eta}_{L_{2,1}}$ and their optimization problems' optimal solutions $\widetilde{\mathbf{x}}^*_{\rho,\eta,L_1}(\mathbf{x})$ and $\widetilde{\mathbf{x}}^*_{\rho,\eta,L_{2,1}}(\mathbf{x})$. To parallel Theorem 1's part 2 for L_1-Moreau envelope, the followings hold*

$$\begin{aligned}\mathrm{ST}_{\rho\eta}\big(-\rho\nabla g^{\rho,\eta}_{L_1}(\mathbf{x})\big) \;&=\; \widetilde{\mathbf{x}}^*_{\rho,\eta,L_1}(\mathbf{x})-\mathbf{x},\\ \mathrm{GST}_{\rho\eta}\big(-\rho\nabla g^{\rho,\eta}_{L_{2,1}}(\mathbf{x})\big) \;&=\; \widetilde{\mathbf{x}}^*_{\rho,\eta,L_{2,1}}(\mathbf{x})-\mathbf{x}.\end{aligned}$$

Proof. We defer the proof to the Appendix. □

Corollary 2. *Suppose that the prediction score function $f_{\mathbf{w},c}$ is λ-weakly convex. Assuming that $0<\rho<\frac{1}{\lambda}$, the Sparse MoreauGrad S-MG$_{\rho,\eta}$ and Group-Sparse MoreauGrad GS-MG$_{\rho,\eta}$ interpretations will be robust to every norm-bounded perturbation $\|\boldsymbol{\delta}\|_2\le\epsilon$ as:*

$$\big\|\textit{S-MG}_{\rho.\eta}(\mathbf{x}+\boldsymbol{\delta})-\textit{S-MG}_{\rho,\eta}(\mathbf{x})\big\|_2 \;\le\; \frac{\epsilon}{\min\{\rho,\frac{1}{\lambda}-\rho\}},$$

$$\big\|\textit{GS-MG}_{\rho,\eta}(\mathbf{x}+\boldsymbol{\delta})-\textit{GS-MG}_{\rho,\eta}(\mathbf{x})\big\|_2 \;\le\; \frac{\epsilon}{\min\{\rho,\frac{1}{\lambda}-\rho\}}.$$

```
Algorithm 1 MoreauGrad Interpretation
Input: data $\mathbf{x}$, label $c$, classifier $f_{\mathbf{w}}$, regulization coeff. $\rho$, stepsize $\gamma$, noise std. parameter $\sigma$, number of updates $T$
Initialize $\mathbf{x}^{(0)} = \mathbf{x}$,
for $t = 0,\dots,T$ do
    if Regularized Mode then
        Draw noise vectors $\mathbf{z}_1,\dots,\mathbf{z}_m \sim \mathcal{N}(\mathbf{0},\sigma^2 I_{d\times d})$
        Compute $\mathbf{g}_t = \frac{1}{m}\sum_{i=1}^{m}\nabla f_{\mathbf{w},c}(\mathbf{x}^{(t)}+\mathbf{z}_i)$
    else
        Compute $\mathbf{g}_t = \nabla f_{\mathbf{w},c}(\mathbf{x}^{(t)})$
    end
    Update $\mathbf{x}^{(t+1)} \leftarrow (1-\frac{\gamma}{\rho})\mathbf{x}^{(t)} - \gamma(\mathbf{g}_t - \frac{1}{\rho}\mathbf{x})$
    if Sparse Mode then
        Update $\mathbf{x}^{(t+1)} \leftarrow \mathrm{SoftThreshold}_{\gamma\eta}\big(\mathbf{x}^{(t+1)}-\mathbf{x}\big)+\mathbf{x}$
end
Output $\mathrm{MG}(\mathbf{x}) = \frac{1}{\rho}\big(\mathbf{x}^{(T)}-\mathbf{x}\big)$
```

Based on the above results, we propose applying the proximal gradient descent algorithm as described in Algorithm 1 to compute the Sparse and Group-Sparse MoreauGrad. We defer discussing Algorithm 1's details to the Appendix.

6. Numerical Results

We conduct several numerical experiments to evaluate the performance of the proposed MoreauGrad. Our designed experiments focus on the smoothness, sparsity, and robustness properties of MoreauGrad interpretation maps as well as the feature maps of several standard baselines. In the following, we first describe the numerical setup in our experiments and then present the obtained numerical results on the qualitative and quantitative performance of interpretation methods.

6.1. Experiment Setup

In our numerical evaluation, we use the following benchmark image datasets: CIFAR-10 [10] consisting of 60,000 labeled samples with 10 different labels, tiny-ImageNet [12] containing 100,000 labels samples with 200 labels, and ImageNet-1K [3] including 1.4 million labeled samples with 1,000 labels. For CIFAR-10 and tiny-ImageNet experiments, we trained a standard ResNet-18 [7] neural network with the softplus activation on the training set. For ImageNet experiments, we used an EfficientNet-b0 network [25] pre-trained on the ImageNet training data. In our experiments, we compared the MoreauGrad schemes with the following baselines: 1) the simple gradient [21], 2) Integrated Gradients [18], 3) DeepLIFT [20], 4) SmoothGrad

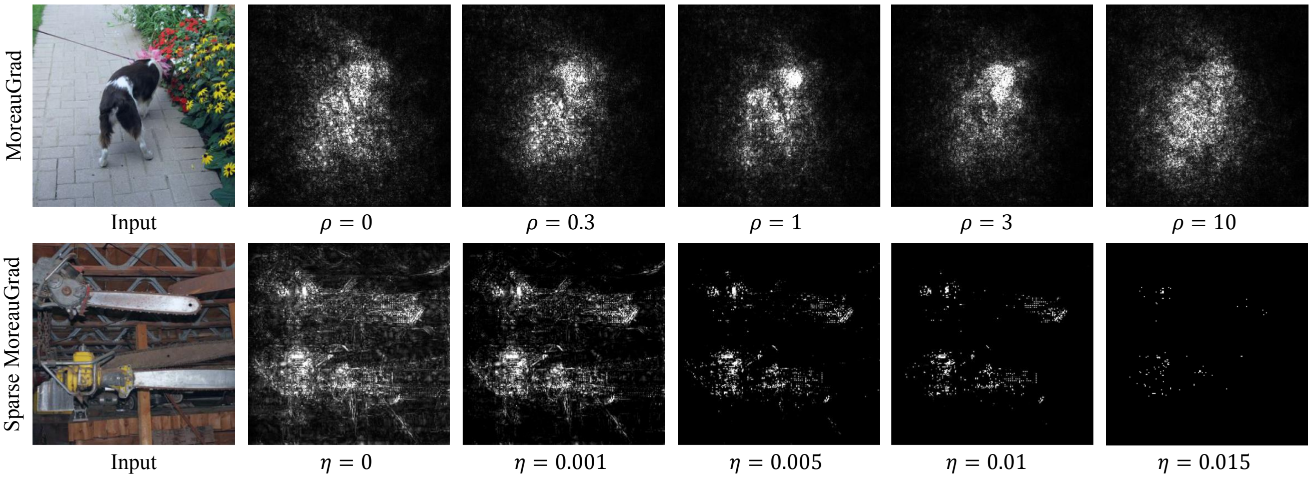

Figure 2. Top: Visualization of MoreauGrad maps with different coefficient ρ's. $\rho = 0$ reduces to the simple gradient method. Bottom: Visualization of Sparse MoreauGrad maps with different coefficient η's. $\eta = 0$ reduces to the Vanilla MoreauGrad.

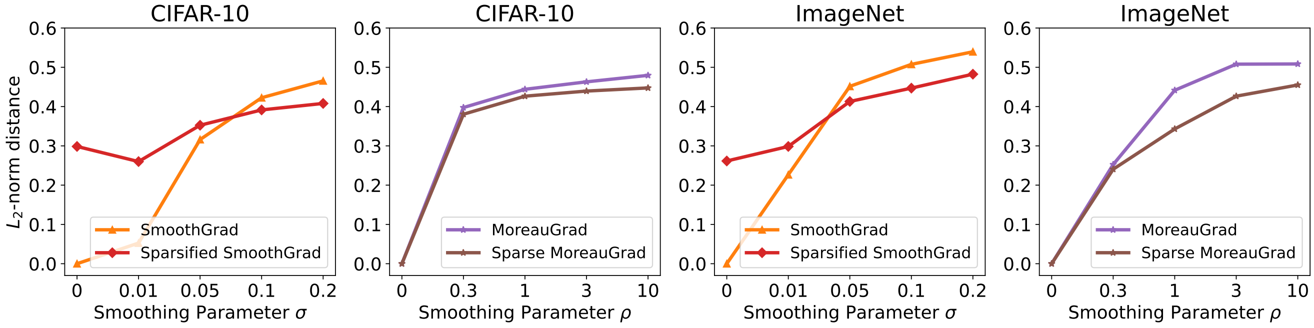

Figure 3. MoreauGrad vs. SmoothGrad gradient discrepancy, measured via the l_2-norm distance from the simple gradient map.

[22], 5) Sparsified SmoothGrad [13], 6) RelEx [14]. We note that for baseline experiments we adopted the official implementations and conducted the experiments with hyperparameters suggested in their work. We present the full implementation details in the Appendix, and our code is available in the supplementary material.

6.2. Effects of Smoothness and Sparsity Parameters

We ran the numerical experiments for unregularized Vanilla MoreauGrad with multiple smoothness coefficient ρ values to show the effect of the Moreau envelope's regularization. Figure 2 visualizes the effect of different ρ on the Vanilla MoreauGrad saliency map. As can be seen in this figure, the saliency map qualitatively improves by increasing the value of ρ from 0 to 1. Please note that for $\rho = 0$, the MoreauGrad simplifies to the simple gradient interpretation. However, as shown in Theorem 1 the proper performance of Vanilla MoreauGrad requires choosing a properly bounded ρ value, which is consistent with our observation that when ρ becomes too large, the Moreau envelope will be computationally difficult to optimize and the quality of interpretation maps could deteriorate to some extent. As numerically verified in both CIFAR-10, tiny-ImageNet, and ImageNet experiments, we used the rule of thumb $\rho = \frac{1}{\sqrt{\mathbb{E}[\|\mathbf{X}\|_2]}}$ measured over the empirical training data to set the value of ρ, which is equal to 1 for the normalized samples in our experiments.

Regarding the sparsity hyperparameter η in Sparse and Group-Sparse MoreauGrad experiments, we ran several experimental tests to properly tune the hyperparameter. Note that a greater coefficient η enforces more strict sparsity or group-sparsity in the MoreauGrad interpretation, and the degree of sparsity could be simply adjusted by changing this coefficient η. As shown in Figure 2, in our experiments with different η coefficients the interpretation map becomes sparser as we increase the L_1-norm penalty coefficient η. Similarly, to achieve a group-sparse interpretation, we used $L_{2,1}$-regularization on groups of adjacent pixels as discussed in Definition 4. The effect of the group-sparsity coefficient was similar to the sparse case in our experiments, as fewer pixel groups took non-zero values and the output interpretations showed more structured interpretation maps when choosing a larger coefficient η. We defer the results on

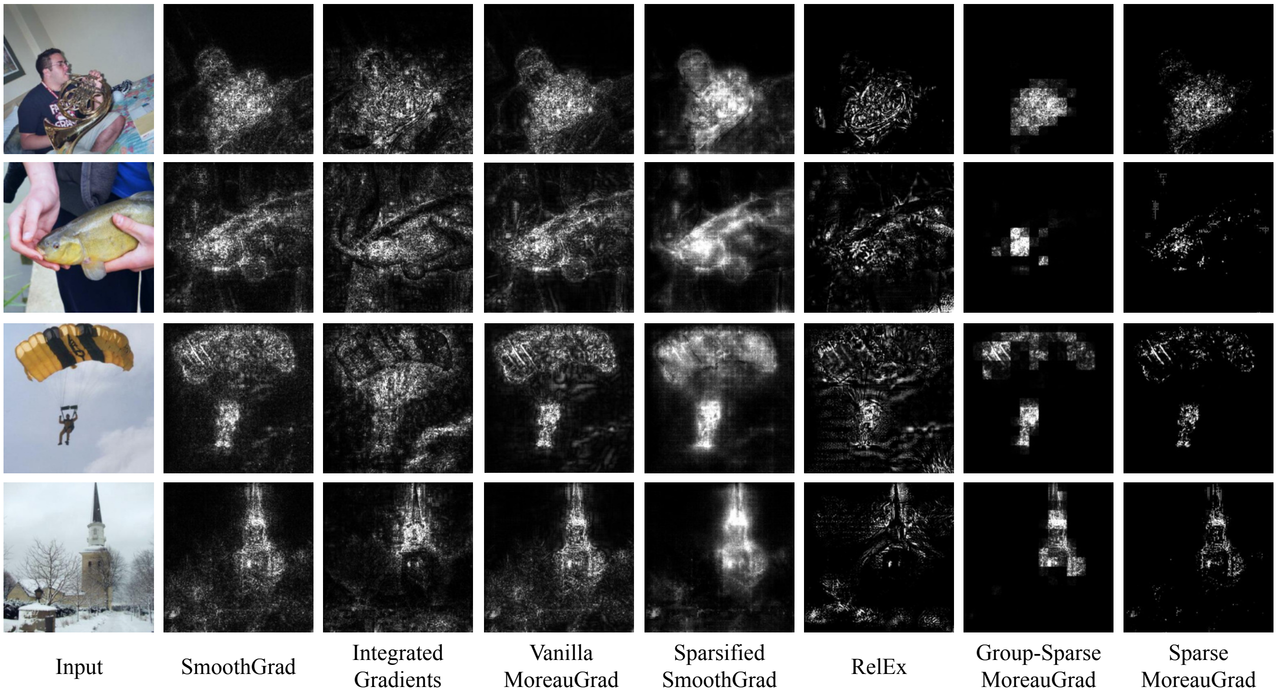

Figure 4. Qualitative comparison between Vanilla, Sparse, Group-Sparse MoreauGrad and the baseline methods.

the group-sparsity hyperparameter effect to the Appendix.

Moreover, to investigate the MoreauGrad's regularization effect, we measured the averaged ℓ_2-norm discrepancy between the MoreauGard and simple gradient maps for different smoothness parameter ρ's. As shown in Figure 3, the ℓ_2-norm discrepancy score grows smoothly with coefficient ρ. We also measured the averaged gradient discrepancy for SmoothGrad methods with different smoothness parameter σ's. We adjusted η in Sparse MoreauGrad to standardize the sparsity level for fair comparisons with Sparsified SmoothGrad. As Figure 3 shows, SmoothGrad led to similar discrepancy values as in MoreauGrad. Also, Integrated Gradients [24] had a similar averaged gradient discrepancy of 0.34 for CIFAR-10 and 0.40 for ImageNet. As the results suggest, the baselines similarly deviate from the simple gradient maps, which could be linked to improved visual quality and robustness, as also observed by [22, 13].

6.3. Qualitative Comparison of MoreauGrad vs. Standard Gradient-based Baselines

In Figure 4, we illustrate the Vanilla, Sparse, and Group-Sparse MoreauGrad interpretation outputs as well as the saliency maps generated by the gradient-based baselines. The results demonstrate that MoreauGrad generates qualitatively sharp and, in the case of Sparse and Group-Sparse MoreauGrad, sparse interpretation maps. As shown in Figure 4, promoting sparsity in the MoreauGrad interpretation maps has improved the visual quality, and managed to erase the less relevant pixels like the background ones. Also, the Group-Sparse MoreauGrad maps successfully exhibit both sparsity and connectivity of selected pixels.

6.4. Robustness of Interpretation Maps

We qualitatively and quantitatively evaluated the robustness of MoreauGrad interpretation. To assess the empirical robustness of interpretation methods, we adopt a L_2-bounded interpretation attack method defined by [13]. Also, for quantifying the empirical robustness, we adopt three robustness metrics. The first metric is the Euclidean distance of the normalized interpretations before and after the attack:

$$D(I(\mathbf{x}), I(\mathbf{x}')) = \left\| \frac{I(\mathbf{x})}{\|I(\mathbf{x})\|_2} - \frac{I(\mathbf{x}')}{\|I(\mathbf{x}')\|_2} \right\|_2 \qquad (6)$$

Note that a larger distance between the normalized maps indicates a smaller similarity and a higher vulnerability of the interpretation method to adversarial attacks.

The second metric is the top-k intersection ratio. This metric is another standard robustness measure used in [6, 13]. This metric measures the ratio of pixels that remain salient after the interpretation attack. A robust interpretation is expected to preserve most of the salient pixels under an attack. The third metric is the structural similarity index measure (SSIM) [27]. A larger SSIM value indicates that the two input maps are more perceptively similar.

Using the above metrics, we compared the MoreauGrad schemes with the baseline methods. As qualitatively shown

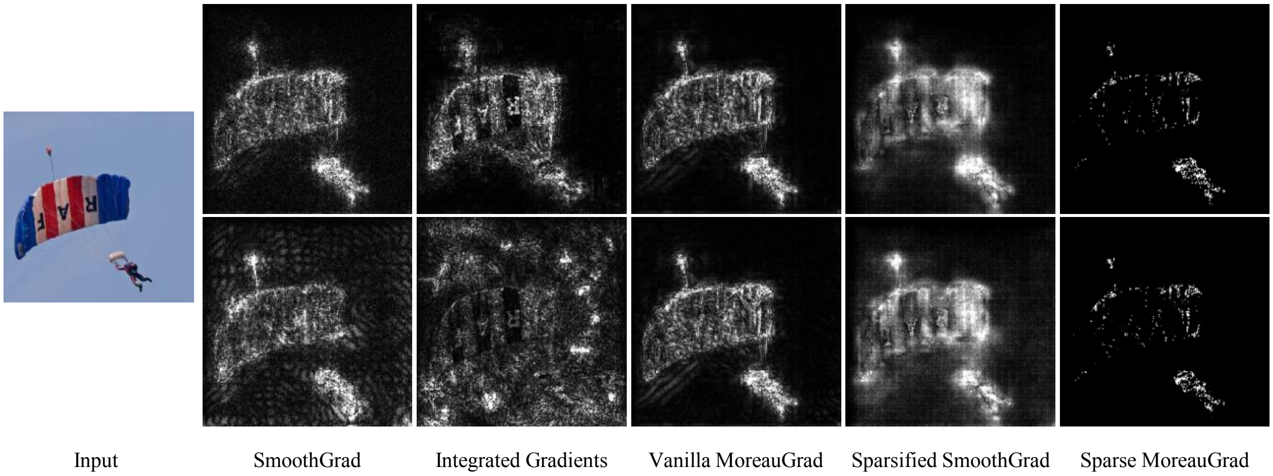

Figure 5. Visualization of robustness against interpretation attacks. The top and bottom rows show original and attacked maps.

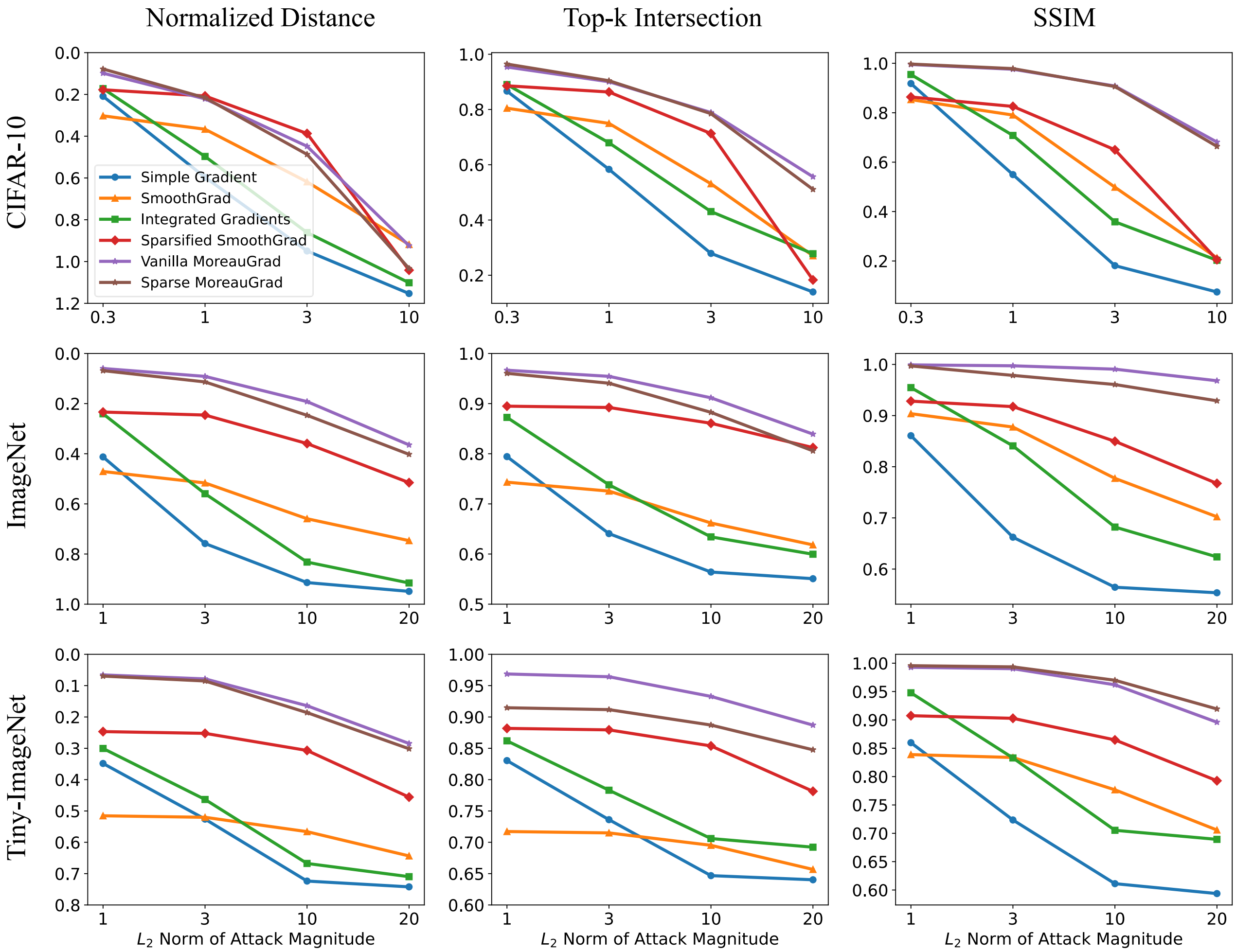

Figure 6. Quantitative robustness comparison between MoreauGrad and the baselines.

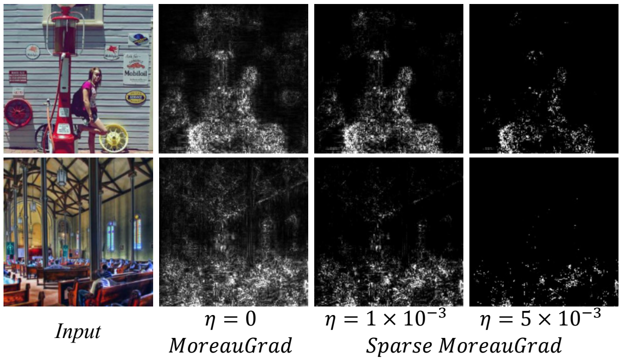

Figure 7. Applications of vanilla and sparse MoreauGrad to explain two misclassifications on the ImageNet test set.

in Figure 5, using the same attack magnitude, the MoreauGrad interpretations are mostly similar before and after the norm-bounded attack. The qualitative robustness of MoreauGrad seems satisfactory compared to the baseline methods. Finally, Figure 6 presents a quantitative comparison of the robustness measures for the baselines and proposed MoreauGrad on CIFAR-10, tiny ImageNet, and ImageNet datasets. As shown by these measures, MoreauGrad outperforms the baselines in terms of the discussed robustness metrics.

6.5. Applications of MoreauGrad Interpretation

As shown in Figure 2, MoreauGrad interpretation maps may be further refined to promote sparsity and retain only the salient pixels. The sparsity level could be flexibly adjusted via parameter η in accordance with the target dataset. Moreover, both vanilla and sparse MoreauGrad can be employed for explaining DNN's misclassifications in image recognition tasks. Figure 7 presents two examples of such misclassifications, where "gas pump" and "church" are erroneously classified as "tricycle" and "restaurant", respectively. Sparse MoreauGrad identifies the source of these errors and relates them to the presence of "a person in the middle of two wheels" and "people sitting on benches", respectively. In the Appendix, we present other examples of using MoreauGrad for explaining DNNs' misclassification.

7. Conclusion

In this work, we introduced MoreauGrad as an optimization-based interpretation method for deep neural networks. We demonstrated that MoreauGrad can be flexibly combined with L_1-regularization methods to output sparse and group-sparse interpretations. We further showed that the MoreauGrad output will enjoy robustness against input perturbations. While our analysis focuses on the sparsity and robustness of the MoreauGrad explanation, studying the consistency and transferability of MoreauGrad interpretations is an interesting future direction. Moreover, the application of MoreauGrad to convex and norm-regularized neural nets could be another topic for future study. Finally, our analysis of ℓ_1-norm-based Moreau envelope could find independent applications in other deep learning contexts.

Acknowledgments

This work was partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China, Project 14209920, and was partially supported by a CUHK Direct Research Grant. Also, the authors would like to thank the anonymous reviewers for their constructive feedback and suggestions.

References

[1] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018. 2

[2] Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. *Advances in neural information processing systems*, 30, 2017. 2, 3

[3] 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. 5

[4] Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[5] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE international conference on computer vision*, pages 3429–3437, 2017. 2, 3

[6] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3681–3688, 2019. 2, 3, 7

[7] 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

[8] Juyeon Heo, Sunghwan Joo, and Taesup Moon. Fooling neural network interpretations via adversarial model manipulation. *Advances in Neural Information Processing Systems*, 32, 2019. 2, 3

[9] Maksims Ivanovs, Roberts Kadikis, and Kaspars Ozols. Perturbation-based methods for explaining

deep neural networks: A survey. *Pattern Recognition Letters*, 150:228–234, 2021. 3

[10] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 1

[12] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. 5

[13] Alexander Levine, Sahil Singla, and Soheil Feizi. Certifiably robust interpretation in deep learning. *arXiv preprint arXiv:1905.12105*, 2019. 3, 6, 7

[14] Dohun Lim, Hyeonseok Lee, and Sungchan Kim. Building reliable explanations of unreliable neural networks: locally smoothing perspective of model interpretation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6468–6477, 2021. 2, 3, 6

[15] 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. 3

[16] Lukas Meier, Sara Van De Geer, and Peter Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008. 2

[17] Sylvestre-Alvise Rebuffi, Ruth Fong, Xu Ji, and Andrea Vedaldi. There and back again: Revisiting backpropagation saliency methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8839–8848, 2020. 2

[18] 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, 5

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

[20] 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, 5

[21] 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, 5

[22] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017. 2, 6, 7

[23] Akshayvarun Subramanya, Vipin Pillai, and Hamed Pirsiavash. Fooling network interpretation in image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2020–2029, 2019. 3

[24] 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, 7

[25] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 5

[26] Jorg Wagner, Jan Mathias Kohler, Tobias Gindele, Leon Hetzel, Jakob Thaddaus Wiedemer, and Sven Behnke. Interpretable and fine-grained visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9097–9107, 2019. 2, 3

[27] 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

[28] Kaidi Xu, Sijia Liu, Pu Zhao, Pin-Yu Chen, Huan Zhang, Quanfu Fan, Deniz Erdogmus, Yanzhi Wang, and Xue Lin. Structured adversarial attack: Towards general implementation and better interpretability. *arXiv preprint arXiv:1808.01664*, 2018. 3

[29] 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

[30] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 2

[31] 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