Building Reliable Explanations of Unreliable Neural Networks: Locally Smoothing Perspective of Model Interpretation (Supplementary Materials)

Dohun Lim Hyeonseok Lee Sungchan Kim

Division of Computer Science and Engineering, Jeonbuk National University, Korea
{imdohun75,hslee0390,s.kim}@jbnu.ac.kr

Contents

S1 The Proofs **1**

S2 Experimental Setups **3**
S2.1 Implementation Details . . . 3
S2.2 Generation of Adversarial Examples . . . 3

S3 Additional Experimental Results **4**
S3.1 Comparison of Perturbation Distance for the Adversarial Attacks . . . 4
S3.2 Results on Similarity of Saliency Maps . . . 4
S3.3 Metrics for Feature Relevance Evaluation . . . 4
S3.4 Results on the Targeted Attacks . . . 5
S3.5 Results on Feature Relevancy of Saliency Maps . . . 7
S3.6 Extracting Explanations of Arbitrary Classes . . . 9
S3.7 Additional Qualitative Results . . . 10
S3.7.1 The Results of the Untargeted Attack for the Natural ResNet-50 . . . 10
S3.7.2 The Results of the Untargeted Attack against the Robust ResNet-50 . . . 13
S3.7.3 The Results of the Targeted Attacks on the Natural ResNet-50: Structured Attack . . . 16
S3.7.4 The Results of the Targeted Attack against the Natural ResNet-50: Unstructured Attack . . . 19

S1. The Proofs

Theorem 1 (Local explanations with respect to label consistency). *Let $\gamma = \alpha \cdot v$ where $||\gamma|| = \alpha$ and $||v|| = 1$. Let $\mathcal{D} = \{x_i\}$ be a set of data samples where $||x_i - x_0|| \leqq \epsilon$. For a given saliency map m calculated from x_0, it holds that*

$$\alpha \geqq \frac{c}{||m||_1} \cdot \frac{2}{||-g(x_0+\alpha v)+g(x_0)|| + 2||g(x_0)||} \tag{S1}$$

where

$$g(x) = -\nabla L(x,m) = \nabla \log f(m \odot x). \tag{S2}$$

It also holds that the loss function in Eq. (1) with respect to $x_0+\gamma$ is upper-bounded as follows:

$$\begin{aligned} &L(x_0+\gamma, m) \leqq \\ &\alpha||m||_1 \left(\frac{||-g(x_0+\alpha v)+g(x_0)||}{2} + ||g(x_0)|| \right). \end{aligned} \tag{S3}$$

Proof. We begin with the definition of Hessian of the loss function $L(x_0, m)$ with respect to input x at x_0 to learn a saliency m as

$$H = \frac{\nabla L(x_0 + \gamma, m) - \nabla L(x_0, m)}{\gamma} = \frac{\nabla L(x_0 + \alpha v, m) - \nabla L(x_0, m)}{\alpha v} \tag{S4}$$

Then,

$$H\gamma = H\alpha v \approx \nabla L(x_0 + \alpha v, m) - \nabla L(x_0, m) \tag{S5}$$

By substituting $L(m \odot x) = -\log f(m \odot x)$ into Eq. (S5), we have

$$H\gamma = -\nabla \log f(m \odot (x_0 + \alpha v)) + \nabla \log f(m \odot x_0) \tag{S6}$$

$$= -\frac{\partial \log f(m \odot x + \alpha m \odot v)}{\partial (m \odot x)} \odot m \Big|_{x_0 + \alpha v} + \frac{\partial \log f(m \odot x)}{\partial (m \odot x)} \odot m \Big|_{x_0} \tag{S7}$$

We rewrite $g(x)$ in Eq. (6) as

$$g(x) = \frac{\partial \log f(m \odot x)}{\partial (m \odot x)}. \tag{S8}$$

Then, Eq. (S7) is given by

$$H\gamma = \{-g(x_0 + \alpha v) + g(x_0)\} \odot m \tag{S9}$$

Using the Cauchy–Schwarz inequality, for the constraint of Eq. (4), the following holds

$$\nabla L(x_0, m)^T \gamma + \frac{1}{2}\gamma^T H \gamma \leqq ||\nabla L(x_0, m)^T|| \cdot ||\gamma|| + \frac{1}{2}||\gamma|| \cdot ||H\gamma|| \tag{S10}$$

$$\tag{S11}$$

By integrating Eq. (S9) into Eq. (S10),

$$\nabla L(x_0, m)^T \gamma + \frac{1}{2}\gamma^T H \gamma \leqq \alpha ||g(x_0)|| \cdot ||m|| + \frac{\alpha}{2}|| - g(x_0 + \alpha v) + g(x_0)|| \cdot ||m|| \tag{S12}$$

Using $||m|| \leqq ||m||_1$, it holds that

$$\nabla L(x_0, m)^T \gamma + \frac{1}{2}\gamma^T H \gamma \leqq \alpha ||g(x_0)|| \cdot ||m||_1 + \frac{\alpha}{2}|| - g(x_0 + \alpha v) + g(x_0)|| \cdot ||m||_1. \tag{S13}$$

Combining Eq. (4) and Eq. (S13) gives

$$\alpha ||g(x_0)|| \cdot ||m||_1 + \frac{\alpha}{2}|| - g(x_0 + \alpha v) + g(x_0)|| \cdot ||m||_1 \geqq c \tag{S14}$$

By rearranging Eq. (S14), we reach Eq. (5),

$$\alpha \geqq \frac{c}{||m||_1} \cdot \frac{2}{|| - g(x_0 + \alpha v) + g(x_0)|| + 2||g(x_0)||}.$$

Finally, combining Eq. (1) and Eq. (S13) gives Eq. (7) as

$$L(x_0 + \gamma, m) \leqq \alpha ||m||_1 \left(\frac{1}{2}|| - g(x_0 + \alpha v) + g(x_0)|| + ||g(x_0)||\right)$$

□

Theorem 2 (Local explanations with respect to saliency map consistency). *Let $\mathcal{D} = \{x_i\}$ be the vicinity of the input data x_0 such that $||x_i - x_0|| \leqq \epsilon$ where ϵ being a small positive number. Then, distance between the gradients of explanations of x_i and x_0 is lower-bounded as follows:*

$$||\nabla L(x_i, m) - \nabla L(x_0, m)|| \leqq ||m||_1 \cdot || - g(x_0 + \alpha v) + g(x_0)||. \tag{S15}$$

Proof. We begin from the following.

$$\nabla L(x, m) = -\log f(m \odot x) = -\frac{\partial \log f(m \odot x)}{\partial (m \odot x)} \odot m \tag{S16}$$

Based on the requirement on the robustness of a saliency map as described in Assumption 2, we assume that the gradient of a data point x_i can be written using the first-order Taylor expansion at x_0, which is given by

$$\nabla L(x_i, m) = \nabla L(x_0 + \gamma, m) \approx \nabla L(x_0, m) + H\gamma. \tag{S17}$$

Then, we consider distance between the gradients of x_i and x_0 as

$$||\nabla L(x_i, m) - \nabla L(x_0, m)|| = ||H\gamma||. \tag{S18}$$

Similar to the steps we took corresponding to Eq. (S10) in the proof of Theorem 1, the following holds

$$||\nabla L(x_i, m) - \nabla L(x_0, m)|| \leqq ||m||_1 \cdot || - g(x_0 + \alpha v) + g(x_0)||.$$

□

S2. Experimental Setups

S2.1. Implementation Details

For the objective function in Eq. (11), given an input image $x_0 = \{x_{0,i}\} \in \mathbb{R}^d$ as a vector of d pixles, we created a batch of 100 neighboring data points with respect to x_0, $\mathcal{D}$, by adding random noise following the normal distribution, $N(0, \sigma)$, to each pixel $x_{0,i}$, where σ is a standard deviation and $\sigma = 0.1 \times (\max(x_{0,i}) - \min(x_{0,i}))$.

A saliency map m is initialized following the uniform distribution on the interval $[0, 0.01]$. We set the parameters of the objective function as $\lambda_1 = 0.0001$ and $\lambda_2 = 1.0$, respectively. We solve Eq. (11) using the stochastic gradient descent (SGD) for 50 epochs with the learning rate set to 0.001. We denote this baseline by $Ours(50)$ in Figure 3. We used normalized gradient in the optimization process when applying to the SGD. We found that the normalization performed better in terms of the quality of saliency maps and the stability of the optimization.

Computation time. The current implementation of the proposed method takes about 17 seconds to solve the optimization with the baseline setting for an image from ImageNet on ResNet-50 using a single RTX 2080 Ti GPU.

Post-processing of saliency maps. It is required to post-process a saliency map to construct an explanation for a given image by multiplying a saliency map and the image. For the proposed method, we use a saliency as a result of the optimization in Eq. (11) directly with no additional processing of the saliency map. However, we applied two strategies differently according to the previous methods. The first strategy is taken from [5] and applies to SmGrad, IntGrad, SimGrad, and DeepLIFT. This strategy takes the expectation of absolute values of 3-channels, RGB, for each pixel in a saliency map. Then, a final saliency is given by normalizing the expectation for each pixel to the 99$^{\text{th}}$ percentile of high value.

The second strategy applies to GradCAM and RT-Sal, where a saliency map consists of single-channel pixel-wise values. Let $g = \{g_i\}$ be a saliency map that is a result of GradCAM or RT-Sal and, thus, not normalized. Then $m = \{m_i\}$, which is a normalized counterpart that we use to evaluate the explanations, is given by

$$m_i = \frac{g_i - \min(g_i)}{\max(g_i) - \min(g_i)}.$$

S2.2. Generation of Adversarial Examples

Implementing the untargeted attack. We used CleverHans [4] to implement the PGD-based untargeted attack. We primarily follow the procedure in [3]. In particular, we applied 40 iterations of the PGD attack when generating adversarial examples. Because the ResNet model in the experiments takes pre-process images where pixel values are scaled to $[-2.117, +2.639]$, we also changed the step size and the range of ℓ_∞-norm of perturbation for the PGD attack in the literature to 0.01 and $\{0.07, 0.1, 0.3, 1, 2, 4, 8\}$ in this study, respectively.

Implementing the targeted attacks. We used the codes released by [2] and [1] to implement the unstructured and the structured attacks respectively.

For the structured attack, we applied the attack for 1500 iterations to the natural ResNet-50 with a learning rate of 0.0002. We set two prefactor values of 10^{11} and 10^{6} for the terms in the objective function of the attack, which correspond to a saliency map and the accuracy loss, respectively [1].

In the case of the unstructured attack, we applied a *top-k fooling* that aims to generate a false saliency map where the top-k feature importance of the saliency map of an input image is reduced as much as possible. We set k to 1000 in the experiments. We applied the attack for 300 iterations. Due to the different preprocessing setting of the model as in the case of the untargeted attack, we rescaled the step size and the perturbation distance used in [2] accordingly.

S3. Additional Experimental Results

S3.1. Comparison of Perturbation Distance for the Adversarial Attacks

Table S1. Difference of adversarial examples compared to their clean counterparts from ImageNet when applying the PGD-based untargeted attack to the natural ResNet-50 by varying ℓ_∞-norm of perturbation.

ℓ_∞-norm of perturbation	0.07	0.1	0.3	1	2	4	8
Difference in ℓ_2-norm	17.8596	24.3589	67.5621	215.2923	405.0308	680.0658	861.8575

Table S2. Difference of adversarial examples compared to their clean counterparts from ImageNet when applying the targeted attacks to the natural ResNet-50 against each method.

Methods	Difference in ℓ_2-norm	IntGrad	SmGrad	RT-Sal	GradCam
Structured attack	min.	14.951	21.448	13.549	15.345
	avg.	628.346	628.28	628.104	628.123
	max.	1007.269	1005.27	1006.81	1006.802
Unstructured attack	min.	8.609	3.301	3.377	3.02
	avg.	54.487	16.847	15.589	49.815
	max.	170.175	22.06	21.812	169.387

Table S2 and Table S1 show that the perturbation distance caused by the targeted attacks is comparable to the case where the ℓ_∞-norm of perturbation is 4. Because the proposed method is robust in such a level of perturbation against the untargeted attack as shown in the paper (Figure 3), we assume that it is impractical to create adversarial images by applying the targeted attacks against our method.

S3.2. Results on Similarity of Saliency Maps

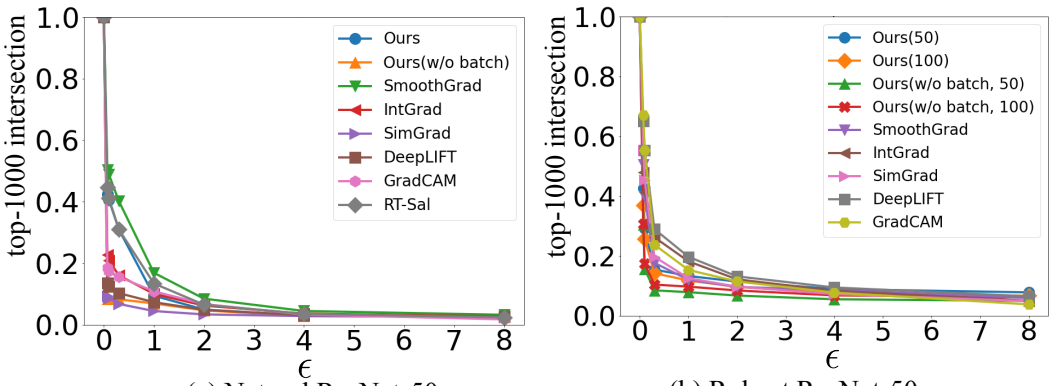
In addition to *Spearman's rank-order correlation* [6], we provide the results of another metric, a *top-k intersection*, to evaluate the similarity of saliency maps. The top-k intersection measures the size of the intersection of the k-most important features between a clean image and its adversary [2].

Figure S1 shows the results of top-1000 intersection, confirming that spatial similarity is unrelated to the fidelity of saliency maps to the model predictions.

S3.3. Metrics for Feature Relevance Evaluation

Deletion and preservation are metrics to evaluate the fidelity of a given saliency map with respect to the softmax score of the corresponding class.

Deletion. A deletion score, which is also known as *pixel flipping*, is calculated as follows. First, we sort pixels of the input image according in descending order of their corresponding values in the saliency map. We apply flipping all pixels to zero in the sorted order, creating a plot that represent a target class score of a given input, which is an explanation in this study. We measure area-under-cover (AUC) of the plot as the deletion score of the saliency map. In general, a significant drop should appear as early as possible with a saliency map of high fidelity. Thus, better deletion results in a low deletion score. See plots on left for each of the methods in Figure S3 and Figure S4.



(a) Natural ResNet-50 (b) Robust ResNet-50

Figure S1. Similarity of the saliency maps of the adversarial examples in the **top-1000 intersection** against the **untargeted attack**.

Preservation. The measurement of a preservation score is similar to the case of a deletion score but pixels of an input image are sorted in the ascending order with respect to the saliency map. Thus, the drop should be as late as possible, meaning that irrelevant pixels are removed earlier than relevant ones. A high score indicates better preservation as opposed to the case of deletion. See plots on the right for each of the methods in Figure S3 and Figure S4.

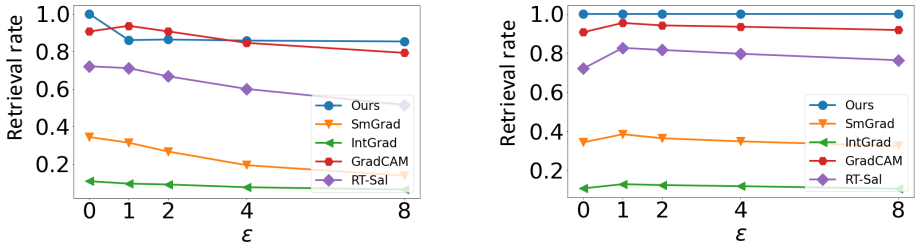
Table S3. Deletion and preservation scores of the methods on ImageNet when applied to the natural ResNet-50.

Method	Deletion (lower is better)	Preservation (higher is better)
SimGrad	0.1336	0.2519
GradCAM	0.1232	0.5647
SmGrad	0.0800	0.3845
IntGrad	0.0907	0.3650
DeepLIFT	0.0980	0.3570
FGVis [8]	0.0644	-
RelEx (proposed)	**0.0567**	**0.4093**

Table S3 shows the deletion and preservation scores by applying each method to *clean images of the ImageNet validation set* on the natural ResNet-50. The score of FGVis [8] is taken from their paper due to the unavailability of implementation, to the best of our knowledge, which was the state-of-the-art deletion score. Our method outperformed FGVis, achieving a new state-of-the-art performance in both deletion and preservation scores.

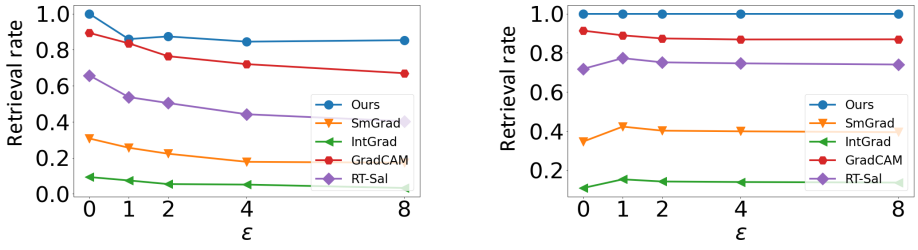
S3.4. Results on the Targeted Attacks

Figure S2 depicts the additional results of the unstructured attack against three more methods, SmGrad, IntGrad, and GradCAM.



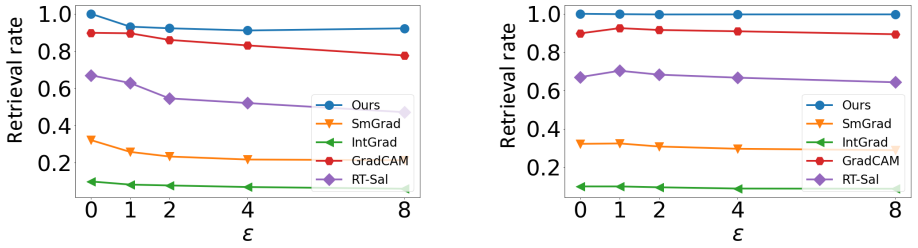
$exp = m_{adv} \odot x_{adv}$ $exp = m_0 \odot x_{adv}$

(a) Attack against SmGrad



$exp = m_{adv} \odot x_{adv}$ $exp = m_0 \odot x_{adv}$

(b) Attack against IntGrad



$exp = m_{adv} \odot x_{adv}$ $exp = m_0 \odot x_{adv}$

(c) Attack against GradCAM

Figure S2. Target class retrieval performance against the **unstructured attacks** on the **natural ResNet-50** for an explanation presented below each plot. Plots correspond to the adversarial images against (a) SmGrad [5], (b) IntGrad [7], and (c) GradCAM [5].

S3.5. Results on Feature Relevancy of Saliency Maps

We provide the deletion and the preservation plots with respect to each of the explanation methods against the untargeted attack.

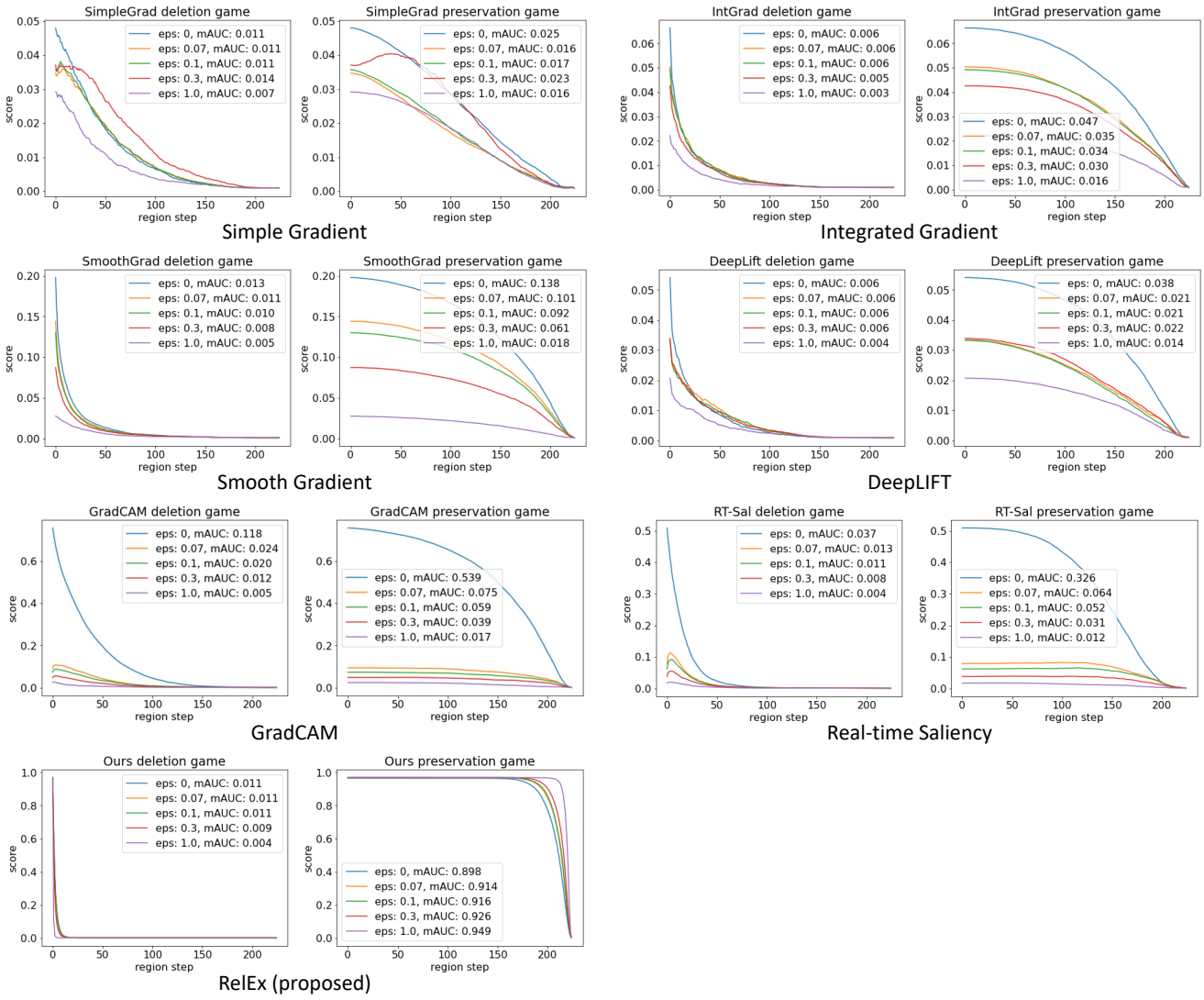


Figure S3. The relevancy of saliency maps for each method in terms of deletion and preservation scores on the **natural ResNet-50** against the **untargeted attack**.

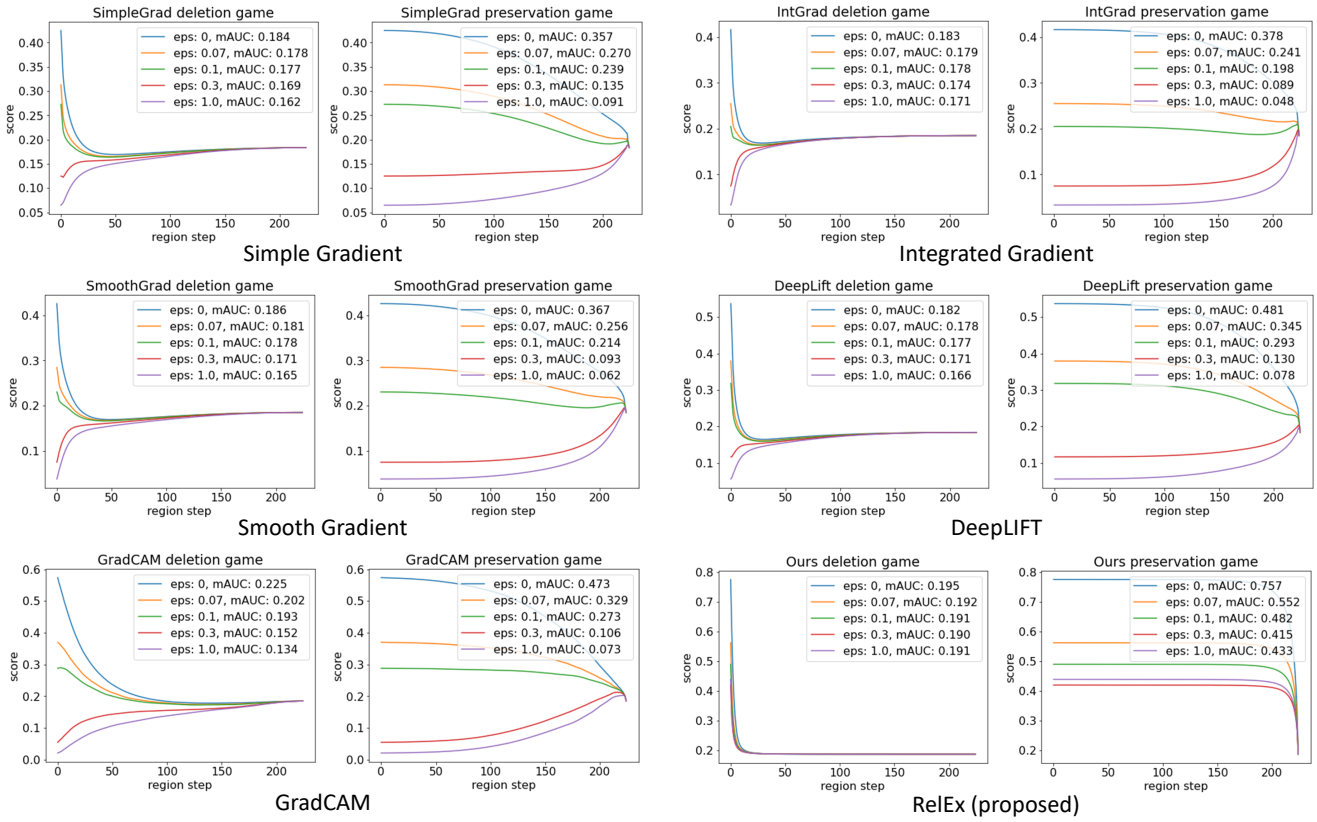
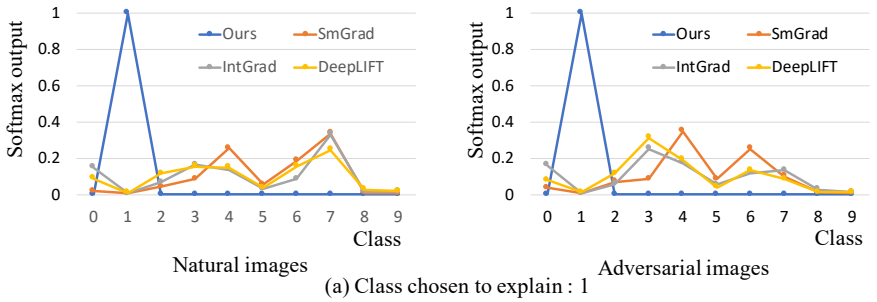


Figure S4. The relevancy of saliency maps for each method in terms of deletion and preservation scores on the **robust ResNet-50** against the **untargeted attack.**

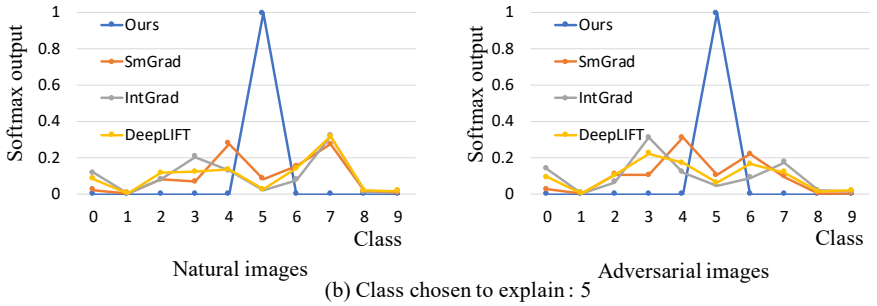
S3.6. Extracting Explanations of Arbitrary Classes

Figure S5 illustrates the additional results of explaining arbitrary classes on CIFAR-10. The proposed method created explanations of three non-target classes, *classes 1, 5*, and *9*, that were applied to randomly selected images of the target *class 7*. The experiments were applied to both the clean images and their adversaries created by the PGD-based untargeted attack.

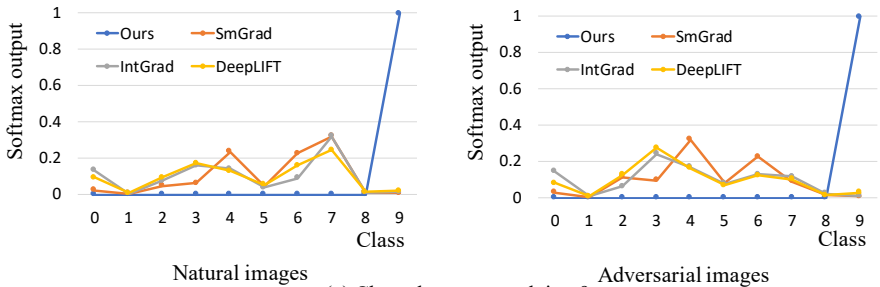
The plots show that the proposed method extracted the explanations of the chosen non-target class correctly. The explanations faithfully contain the information of their corresponding classes. In other words, almost no evidence on other classes were captures in the explanations.



(a) Class chosen to explain : 1



(b) Class chosen to explain : 5



(c) Class chosen to explain : 9

Figure S5. Explaining non-target classes that are arbitrarily chosen. The horizontal axes of the plots enumerate all classes in the **CIFAR-10 dataset** on the **natural ResNet-18** and the vertical ones correspond to the softmax scores of all the classes for a given explanation of the arbitrary classes.

S3.7. Additional Qualitative Results

We present additional qualitative results of saliency maps. In the following figures, numbers below images represent the softmax scores of the target classes of the images. Numbers below saliency maps represent the softmax scores of the target classes with respect to explanations corresponding to the saliency maps. ϵ denotes ℓ_∞-norm of perturbation. Saliency maps are best viewed zoomed-in on screen.

S3.7.1 The Results of the Untargeted Attack for the Natural ResNet-50

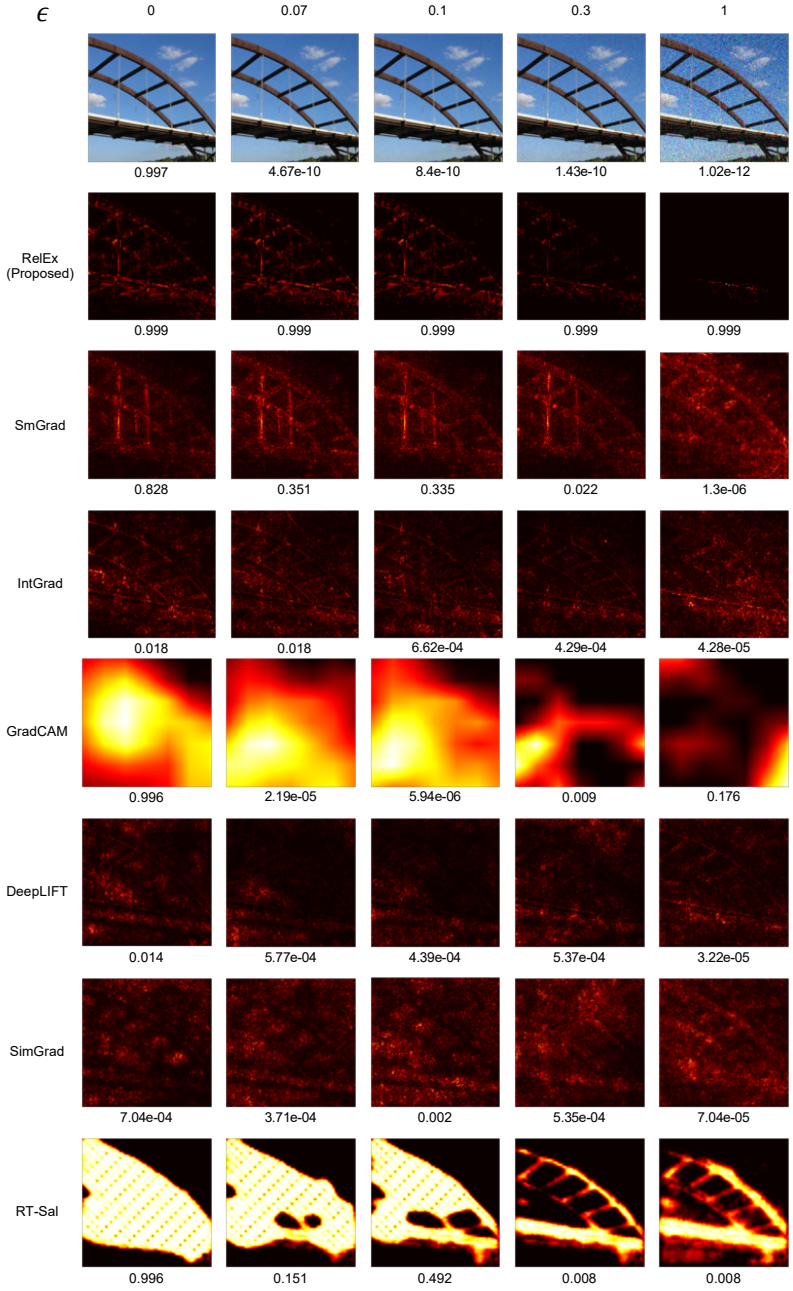


Figure S6. Qualitative results of the methods on the **natural ResNet-50** against the **untargeted attack**.

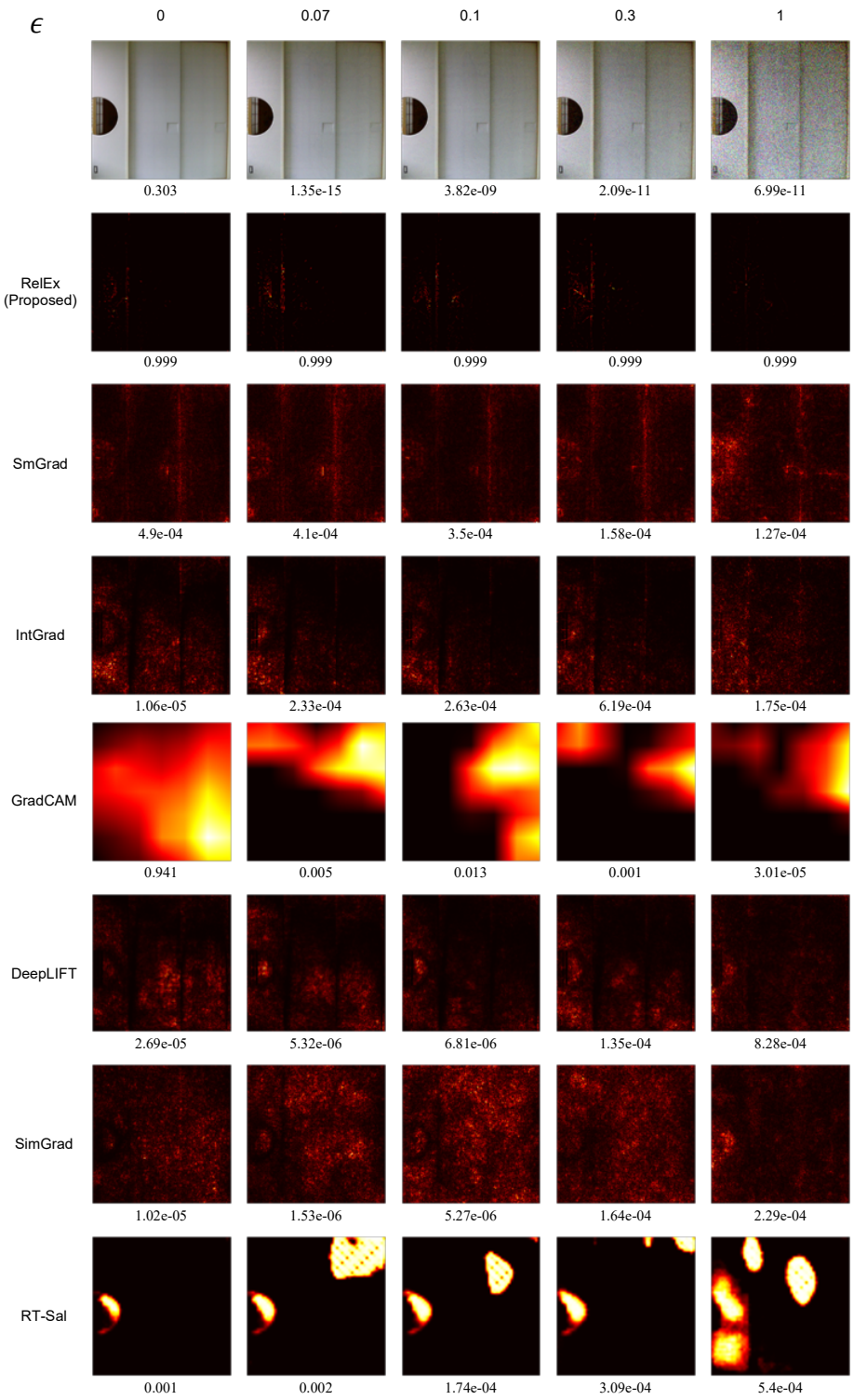


Figure S6a. Qualitative results of the methods on the **natural ResNet-50** against the **untargeted attack** (continued).

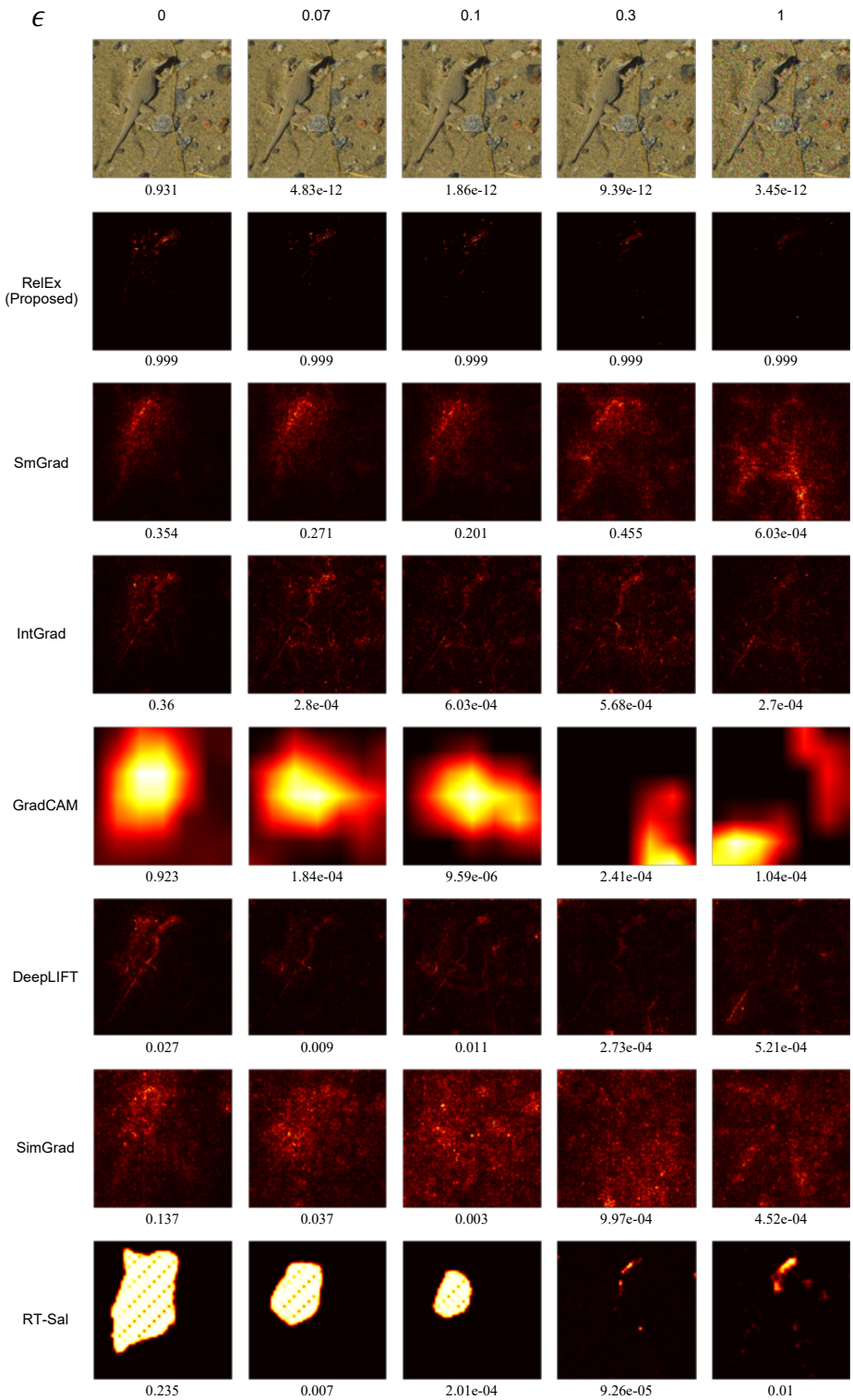


Figure S6b. Qualitative results of the methods on the **natural ResNet-50** against the **untargeted attack** (continued).

S3.7.2 The Results of the Untargeted Attack against the Robust ResNet-50

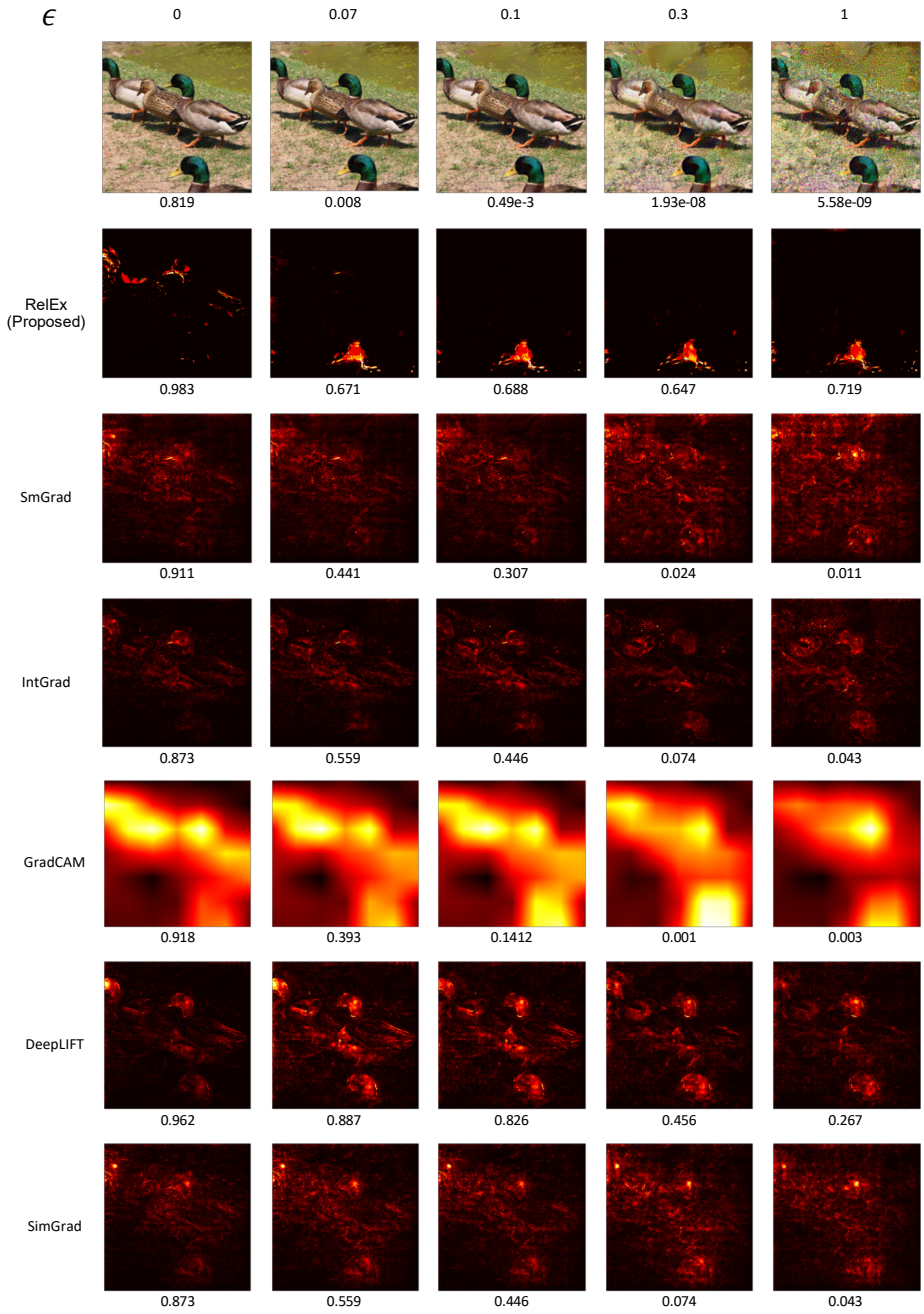


Figure S7. Qualitative results of the methods on the **robust ResNet-50** against the **untargeted attack**.

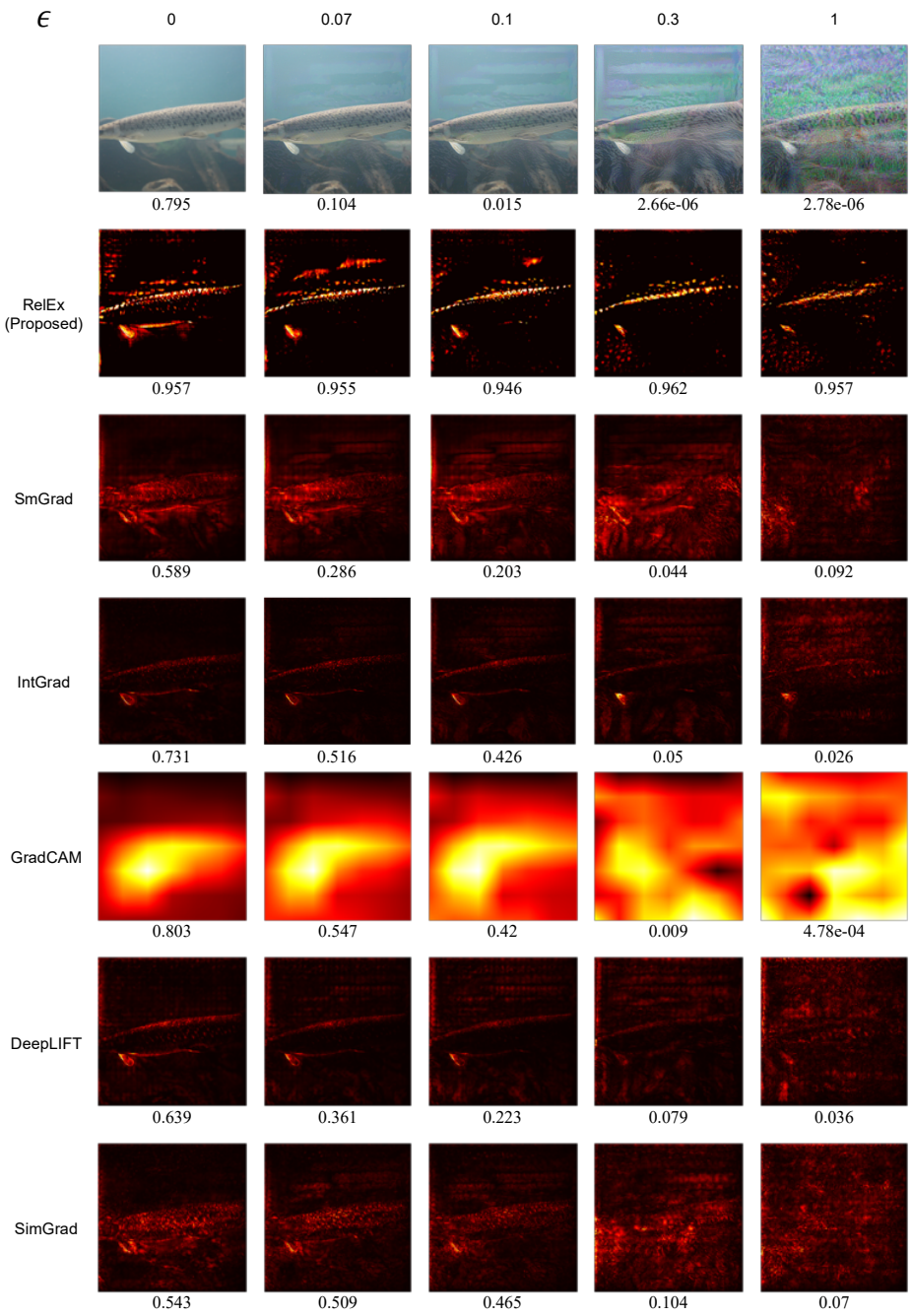


Figure S7a. Qualitative results of the methods on the **robust ResNet-50** against the **untargeted attack** (continued).

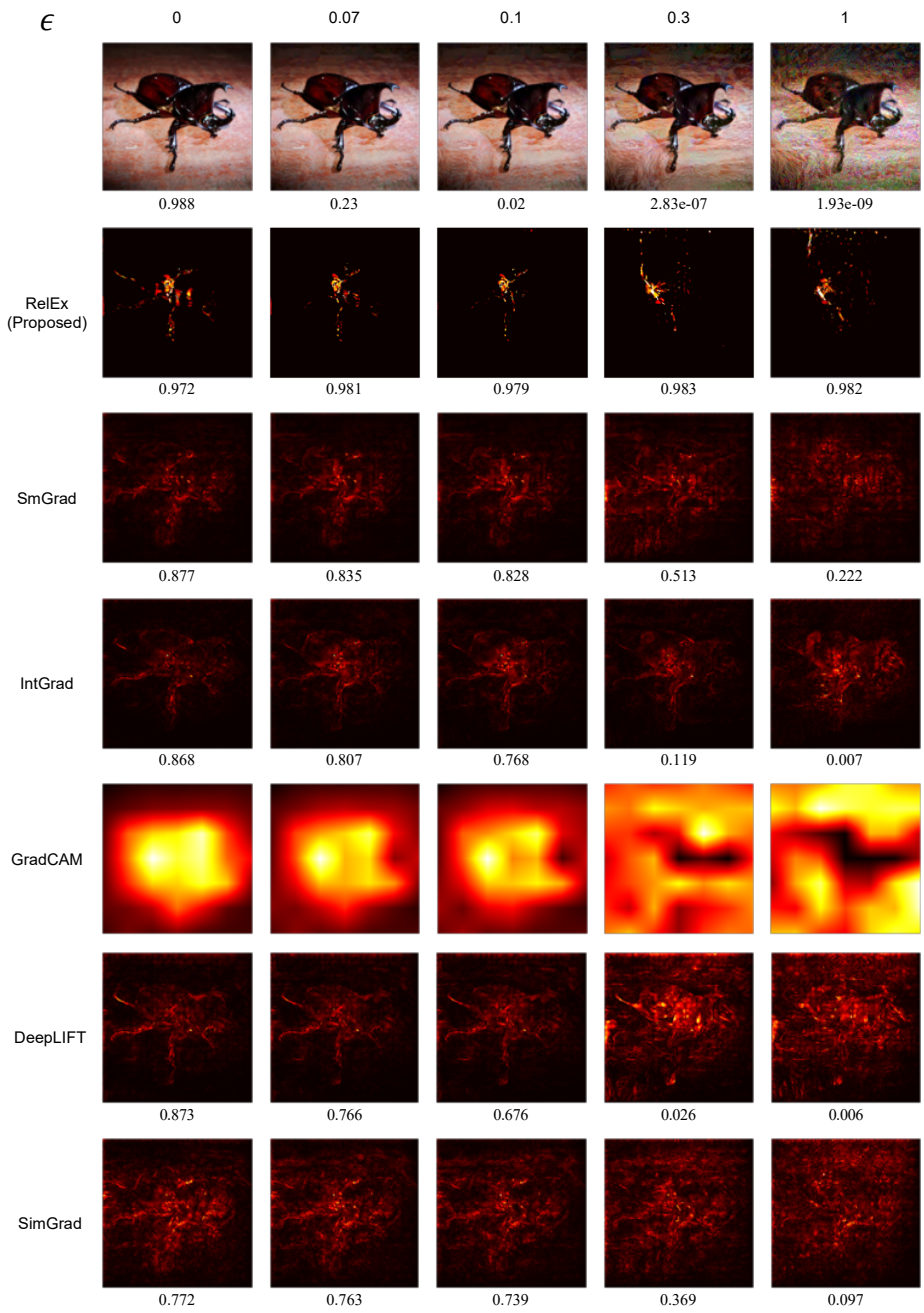


Figure S7b. Qualitative results of the methods on the **robust ResNet-50** against the **untargeted attack** (continued).

S3.7.3 The Results of the Targeted Attacks on the Natural ResNet-50: Structured Attack

The structure attack aims to change the saliency of an *original* image to that of the *target* image. We use the "cat" picture as the target image for all examples below. The first and the second row represent saliency maps of the target and the original images by each method, respectively. We applied the structured attack to SmGrad, IntGrad, GradCAM, and RT-Sal (the red box), creating adversarial images for each method being attacked. Then, we extracted explanations of the adversarial images against a method by using other methods (the blue box) as shown below. Our method RelEx generates consistent saliency maps against the attacks to all the methods unlike other methods.

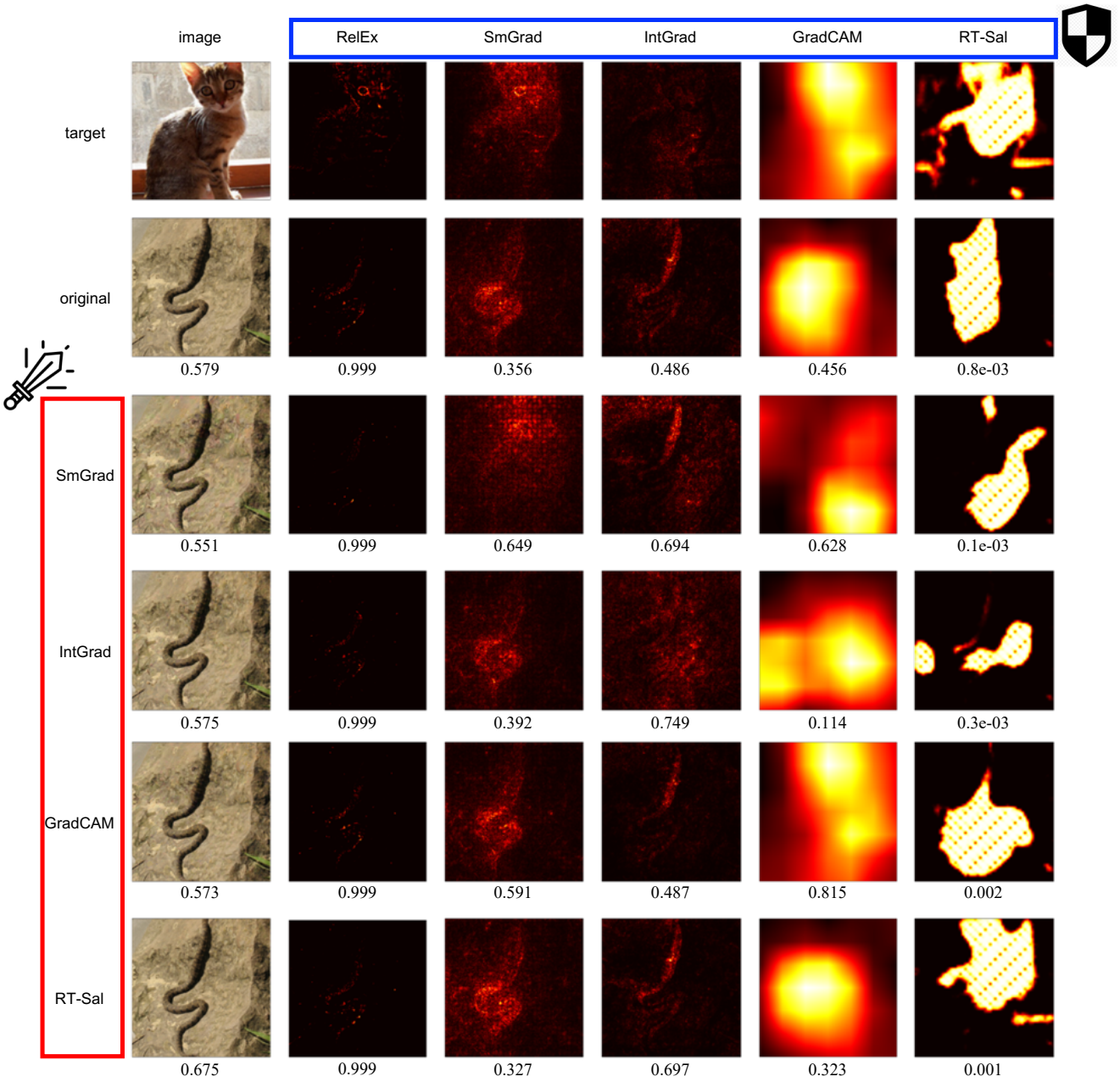


Figure S8. Qualitative results of the methods on the **natural ResNet-50** against the **structured attack**.

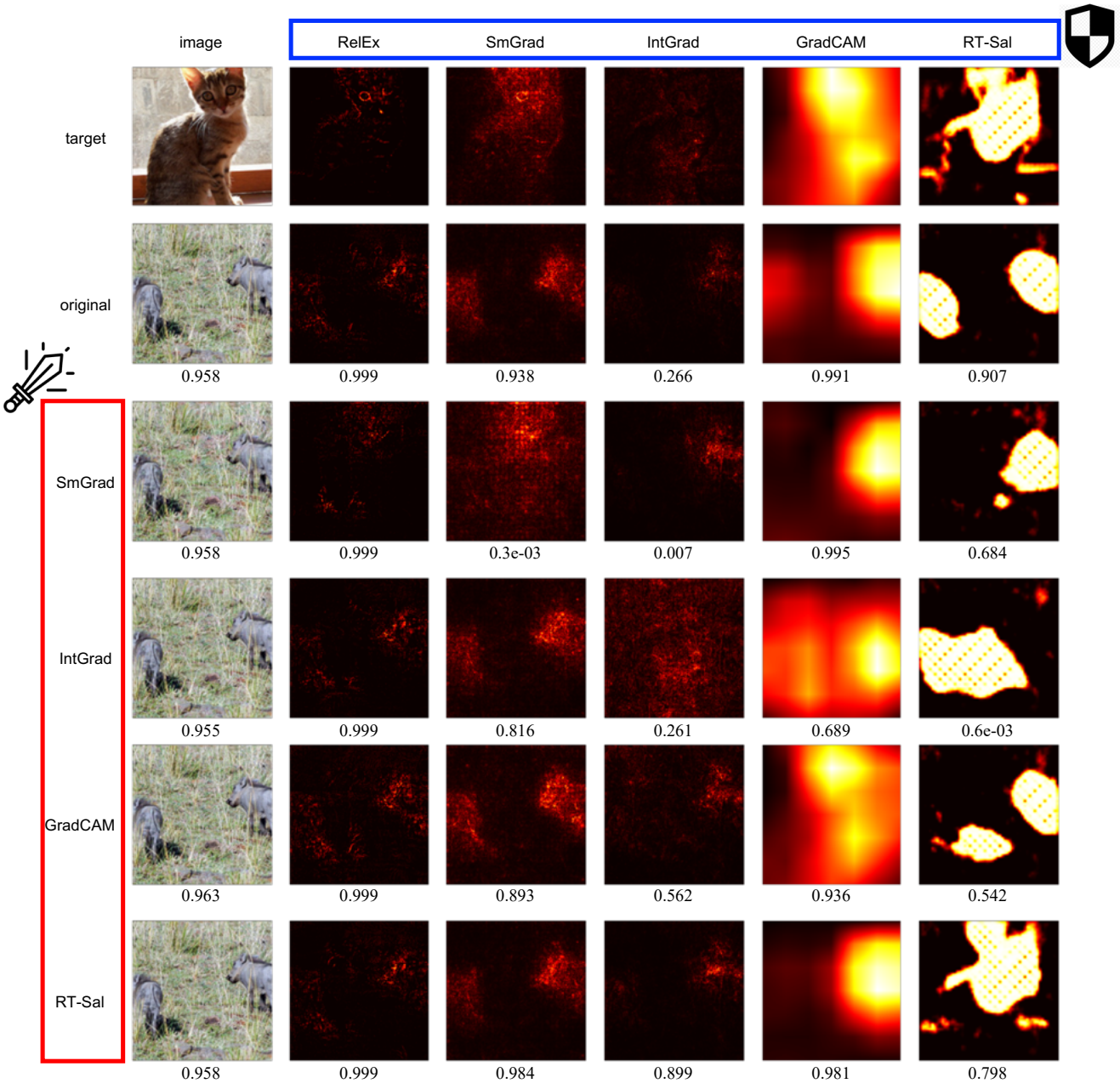


Figure S8a. Qualitative results of the methods on the **natural ResNet-50** against the **structured attack** (continued).

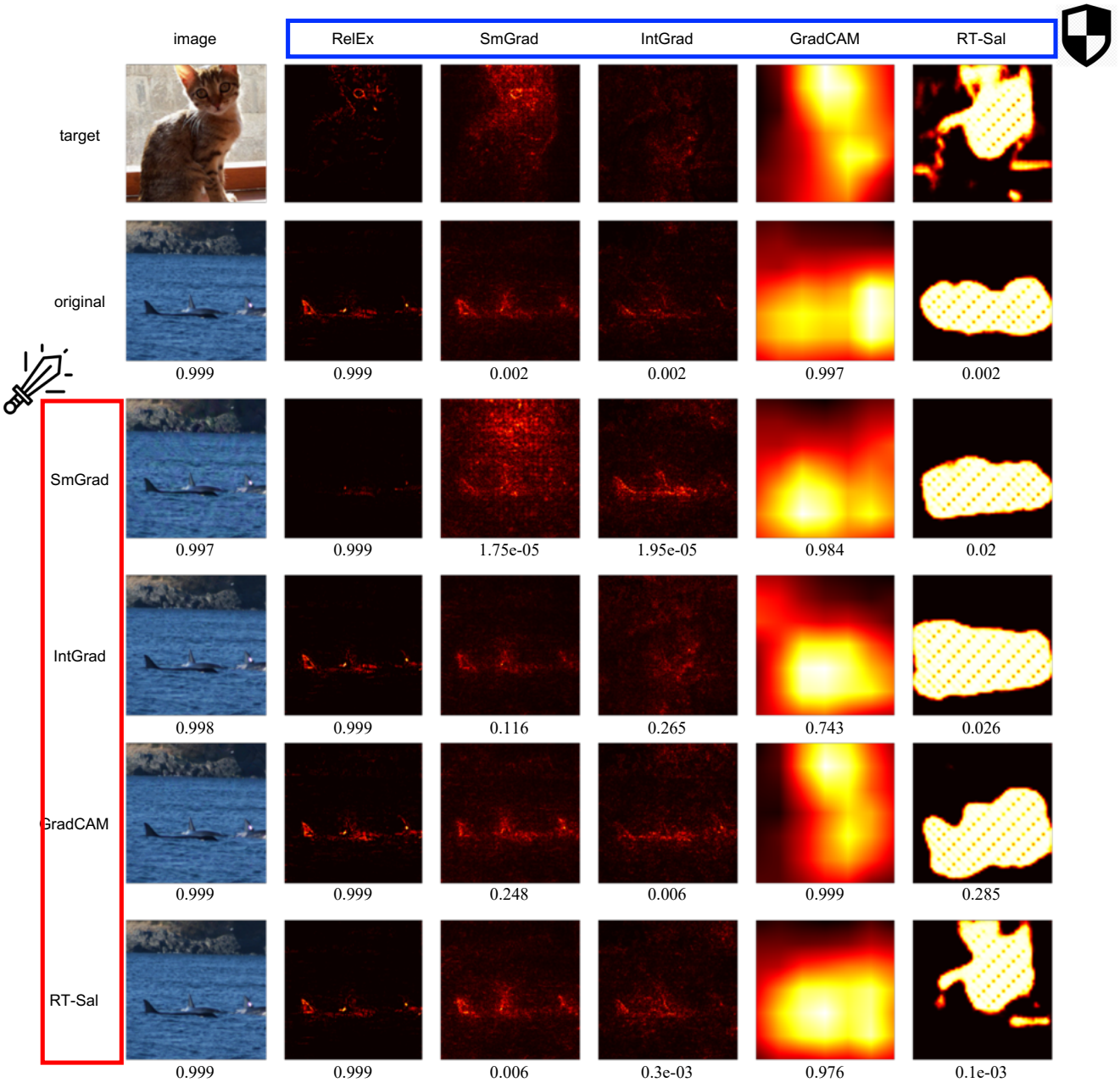


Figure S8b. Qualitative results of the methods on the **natural ResNet-50** against the **structured attack** (continued).

S3.7.4 The Results of the Targeted Attack against the Natural ResNet-50: Unstructured Attack

The unstructure attack aims to change the saliency of an original image to that of the *target* image. We applied the unstructured attack to SmGrad, IntGrad, GradCAM, and RT-Sal (the red box), creating adversarial images for each method being attacked. We chose one method being attacked, of which saliency maps appear to change significantly. Then, we extracted explanations of the adversarial images against the selected method by using other methods (the blue box) as shown below. We present the selected method at the bottom of each figure. Our method RelEx generates consistent saliency maps with high target class scores of corresponding explanations unlike other methods.

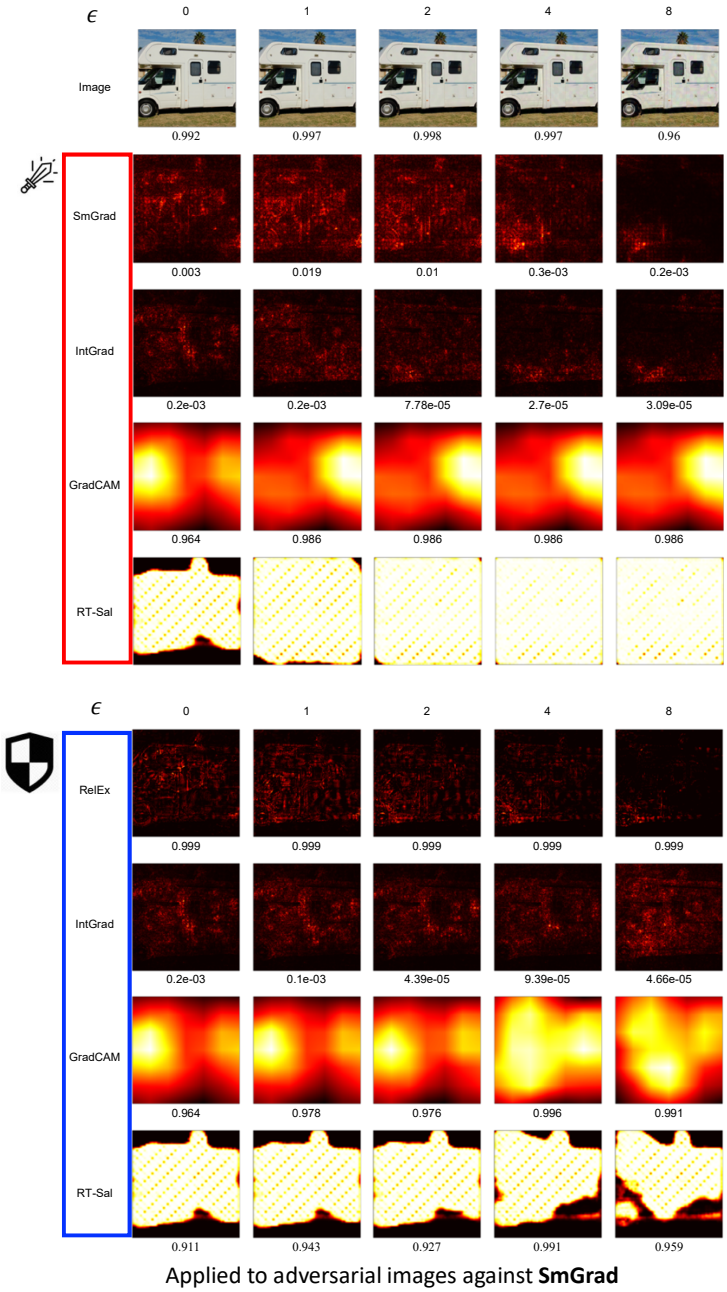


Figure S9. Qualitative results of the methods on the **natural ResNet-50** against the **unstructured attack**.

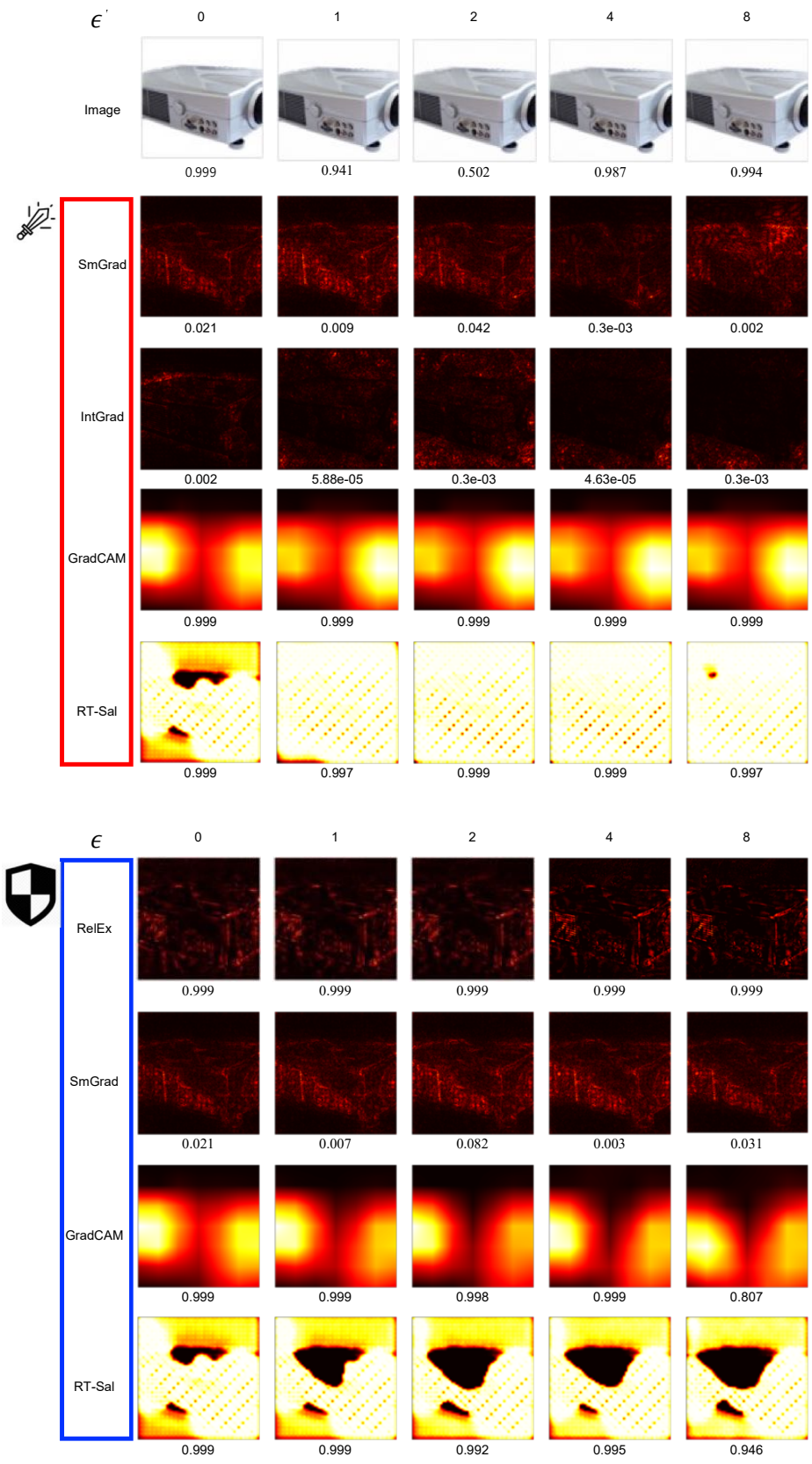


Figure S9a. Qualitative results of the methods on the **natural ResNet-50** against the **unstructured attack** (continued).

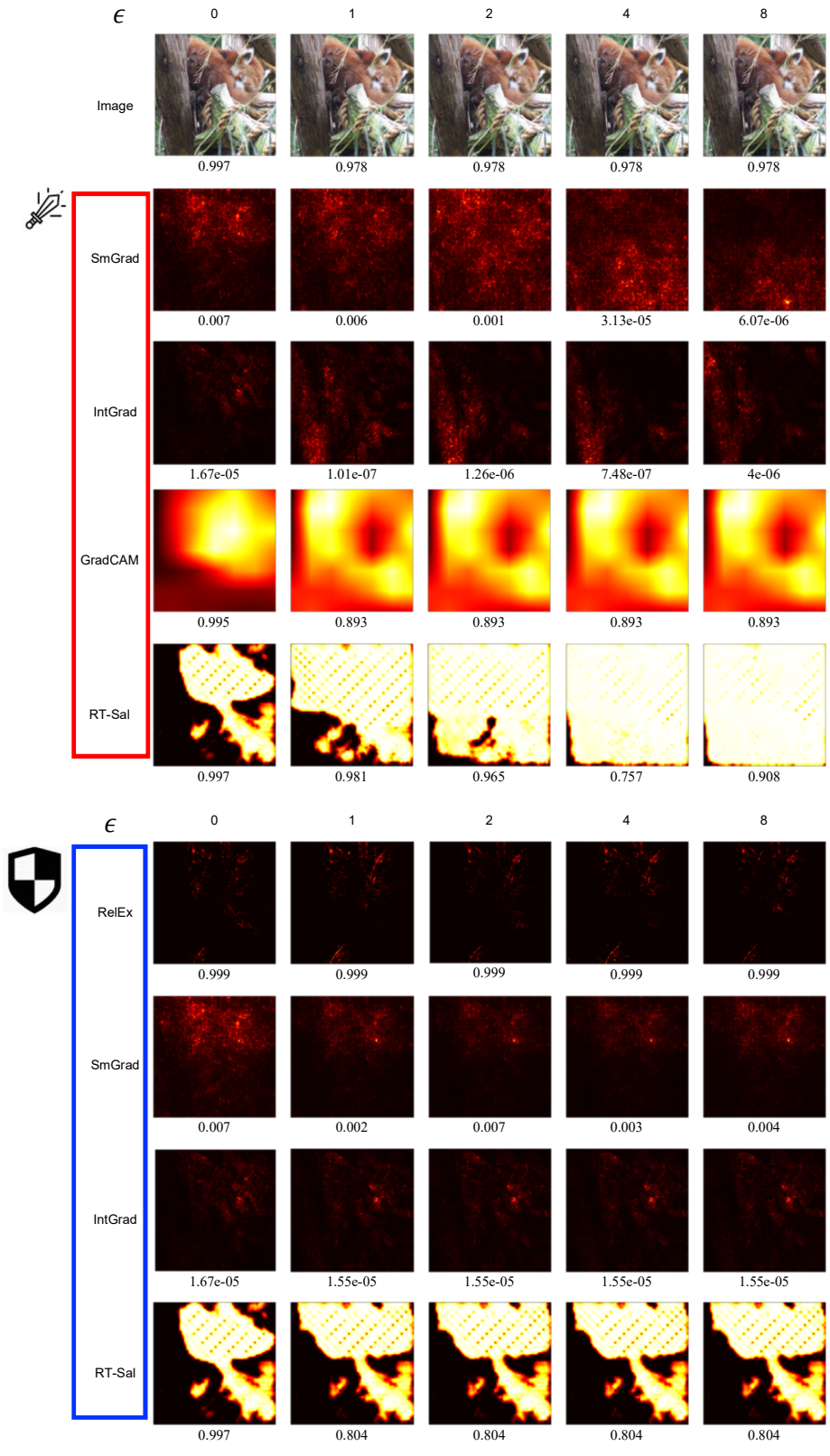


Figure S9b. Qualitative results of the methods on the **natural ResNet-50** against the **unstructured attack** (continued).

References

[1] Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. In *Advances in Neural Information Processing Systems*, pages 13589–13600, 2019. 4

[2] 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. 4

[3] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 3

[4] Nicolas Papernot, Fartash Faghri, Nicholas Carlini, Ian Goodfellow, Reuben Feinman, Alexey Kurakin, Cihang Xie, Yash Sharma, Tom Brown, Aurko Roy, Alexander Matyasko, Vahid Behzadan, Karen Hambardzumyan, Zhishuai Zhang, Yi-Lin Juang, Zhi Li, Ryan Sheatsley, Abhibhav Garg, Jonathan Uesato, Willi Gierke, Yinpeng Dong, David Berthelot, Paul Hendricks, Jonas Rauber, and Rujun Long. Technical report on the cleverhans v2.1.0 adversarial examples library. *arXiv preprint arXiv:1610.00768*, 2018. 3

[5] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017. 3, 6

[6] Charles Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 100(3/4):441–471, 1987. 4

[7] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3319–3328, 2017. 6

[8] 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 Conference on Computer Vision and Pattern Recognition*, pages 9097–9107, 2019. 5