How Does Model Explainability Against Adversarial Attack on Computer Vision?

Jun Huang, Student ID: 40168167
Electrical and Computer Engineering, Concordia University
Email: jun.huang@concordia.ca

Abstract

Deep Learning, a subset of AI, is crucial in industries like healthcare, finance, transportation, and retail, mainly in Computer Vision, where it is used for medical imaging, autonomous vehicles, and more. The lack of interpretability in deep learning is a major challenge, but Explainable AI (XAI) aims to address it by providing techniques such as visualizations, feature importance ranking, and counterfactual explanations to enable humans to understand the decision-making process. On the other hand, Adversarial attacks exploit deep learning models' weaknesses, compromising security in computer vision applications like autonomous driving and biometric authentication, and both defenders and attackers seek to leverage the models' properties. Studying how adversarial attacks can fool Deep Learning models through XAI methods is crucial to achieving adversarial robustness and requires further research. Adversarial Attack is a severe issue and should be counted as an evaluating perspective of XAI methods. This report evaluates how the Adversarial Attack affects the XAI method with our defined metric, particularly for the image classification tasks. We define four metrics by leveraging the ground-truth segmentation on the image as the ground-truth explanation. We perform experiments on eight different Adversarial Attacks and five different XAI methods. We observe that the Guided Backpropagation Gradients are the most vulnerable XAI method against Adversarial Attacks. Nevertheless, XAI methods like SmoothGrad and our developed XAI method, namely, the X Gradients, are robust against all selected attacks. We also find that the Integrated Gradients method has the potential improvement if introducing some of the ideas of the attacks.

Index Terms

eXplainable AI (XAI), Adversarial Attack, Computer Vision, Saliency Explanation, Image Classification

I. Introduction

A branch of Artificial Intelligence (AI) is Deep Learning (DL), and it is taking an essential role in the daily lives of all humankind. Deep learning has found numerous applications in various real-world cases in crucial areas such as healthcare, finance, transportation, natural language processing, and retail. Particularly in the Computer Vision (CV) domain, the development of Deep Learning affects the evolution of a wide range of image and video artificial intelligence applications. In autonomous vehicles, computer vision algorithms are used to identify objects such as pedestrians, other vehicles, and traffic signs. Deep learning algorithms are used in healthcare to develop more accurate medical imaging and diagnosis systems. These algorithms can be trained to detect abnormalities in X-ray, CT, or MRI scans, which can help doctors make more accurate diagnoses and treatment plans.

However, one of the main challenges of deep learning is its need for more interpretability, i.e., the inability to explain how it arrived at a particular decision. This lack of transparency makes it difficult for humans to trust and validate the decisions made by deep learning models. XAI seeks to address this challenge by providing methods and techniques that allow humans to understand the decision-making process of deep learning models. Some popular methods used in XAI include visualizations, feature importance ranking, and counterfactual explanations. These methods enable humans to understand why a deep learning model made a particular decision and what factors influenced the decision.

Despite the tremendous success of the joint research, a significant threat beneath it comes from the security aspect, namely, adversarial attack on Computer Vision. Adversarial attacks pose a significant threat to computer vision deep learning applications by exploiting the model's weaknesses and manipulating it

to misclassify inputs with small perturbations. These attacks can compromise the security of applications that rely on the accuracy of deep learning models, such as autonomous driving, biometric authentication systems, and medical image analysis. Defenders and Attackers try hard to leverage the properties and the mechanism of the Deep Learning model from the training to the testing stage, from the data-driven to the model-driven perspective. Existing research reveals that interpretability and attack robustness have strong relations [1]. There needs to be more research on how many different kinds of adversarial attacks on Computer Vision would fool the explanation generated from the post-hoc XAI methods. Therefore, studying how adversarial attack fools the Deep Learning model by explainable AI is a promising direction to achieve adversarial robustness.

This report intends to provide an initial study on how state-of-the-art adversarial attacks affect the explanation generated by XAI methods. More specifically, we study the loss between the saliency explanation of the original image and adversarial images' explanations. We utilize the ground-truth segmentation of the dataset as the ground-truth saliency explanation and define four metrics to evaluate: (1) How much does an adversarial sample disturb the explanation method? (2) Is it a positive or negative effect? We hope this initial study could be a novel angle of Adversarial Attack defense solutions or even turn around to benefit the development of the XAI methods.

II. Literature Review

In this section, we first introduce the background of the Adversarial Attack and how it is applied to Computer Vision related tasks. We then present the current research state of the explainable AI for Computer Vision. Finally, we brief the existing works on the joint topic of the Adversarial Attack and explanation XAI techniques.

A. Adversarial Attack over Computer Vision

A survey in 2018 [2] establishes the catalog of state-of-the-art Adversarial Attack techniques and research projects in the Deep Learning domain. The survey reviews the existing approaches for adversarial attacking and possible defense solutions for the attacks. It introduces fifteen different attacking techniques targeting the input data, including: (1) L-BFGS Attack [3], (2) Fast Gradient Sign Method (FGSM) [3], (3) Basic Iterative Method (BIM) [4], (4) Jacobian-based Saliency Map Attack (JSMA) [5], and (5) Carlini and Wagner's Attack (C&W) [6]. The general idea of the input data attack is to perturb the input image in an imperceptible human range, while such perturbation is well enough to fool the model's prediction. A good example is the Fast Gradient Sign Method (FGSM) attack. The perturbation is generated by:

$$\eta = \epsilon sign(\nabla_x J_\theta(x, l)) \tag{1}$$

where ϵ is the magnitude of the perturbation, $J_\theta(x, l)$ is the prediction function of input x w.r.t the label l. The perturbation stems from the gradient of the prediction and is generalized with the $sign(\cdot)$ function. After that, the adversarial sample is generated by $x' = x + \eta$. Fig. 1 shows the human imperceptible perturbation by the FGSM attack and how much it fools the model prediction.

Another study in 2021 [7] further formalized the Adversarial Attack on Computer Vision. The study provides a list of well-defined terms in the joint topics and further introduces more advanced attacks, including: (1) Projected Gradient Descent (PGD) [8], (2) Guided Adversarial Margin Attack (GAMA) [9], (3) Customized Adversarial Boundary (CAB) [10], and (4) a wide range of enhanced attacks of the original FGSM and PGD attacks. The study further discusses how adversarial study in Deep Learning can benefit other perspectives, such as (1) improving the model's performance [11]. (2) reprogram the target model [12] and (3) reducing bias of the model [13]. Moreover, it explores the linkage between the attacks and model interpretation [14], a topic that stems from the explainable AI domain.

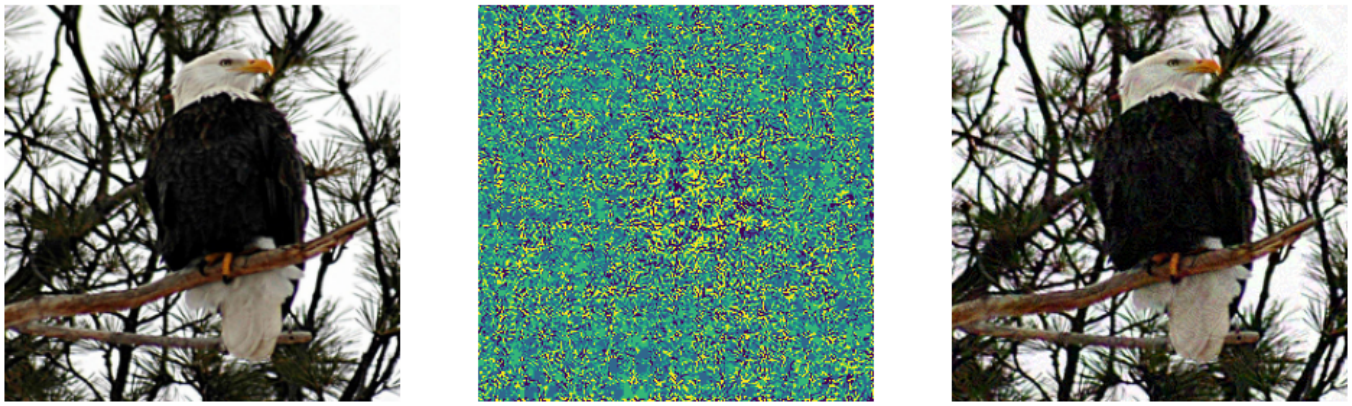

Fig. 1: Left: original image with 99.86% prediction confidence on label “Bald Eagle”. Middle: Perturbation generated by FGSM with 0.005 ϵ. Right: Perturbed adversarial image with 75.97% prediction confidence on label “Hornbill” and 23.1% prediction confidence on label “Bald Eagle”.

B. XAI Techniques in Computer Vision

The study of XAI in Computer Vision domains mainly explains the saliency information of the classified input images. To our best knowledge, two main branches can be concluded for the existing saliency XAI methods. A branch is called Perturbation-based: Methods like SHAP [15] and LIME [16] manipulate parts of the image to generate explanations (model-agnostic). Those methods generate the saliency explanation by perturbing a certain region of pixels to calculate the importance or contribution of these pixels. The perturbation requires a large amount of computation effort. Hence, it is expensive to calculate the saliency value pixel by pixel. A flourishing branch is the gradient-based XAI method. Several techniques calculate

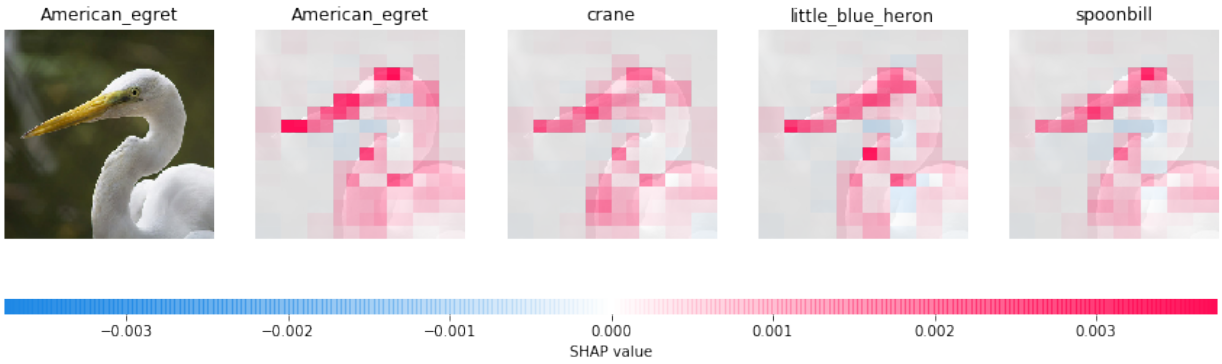

Fig. 2: Saliency explanation generated by SHAP method for an ImageNet sample.

the gradient of the prediction or classification score in relation to the input features. The various gradient-based approaches differ primarily in their methods for computing the gradient. The Gradients or Vanilla Gradient Saliency Map is the classic method that utilizes the gradient of the output activation. Suppose we have:

$$M_c(x, \eta_i) = \nabla f_c(x + \eta_i) = \frac{\partial f_c(x + \eta_i)}{\partial (x + \eta_i)} \tag{2}$$

where $f_c(x)$ is the activation function regarding class c of the input x, and η is the noise. We list the non-intrusive gradient-based method as the following.

- **Vanilla Gradient Saliency Map [17]**:

$$M_c(x, 0) = \nabla f_c(x) \tag{3}$$

- Suppose we have $M_c^p(x) = [M_c(x, \eta_1), M_c(x, \eta_2), \ldots, M_c(x, \eta_p)]$. Then the **classical SmoothGrad (SG) [18]** is formalized by:

$$\overline{M_c^p(x)} = \frac{1}{p} \sum_{i=1}^{p} M_c(x, \eta_i) \tag{4}$$

  where p is the number of iterations, and η_i is generated from the Gaussian distribution with a mean value of 0.
- **VarGrad [19]**:

$$Var(M_c^p(x)) = \frac{1}{p} \sum_{i=1}^{p} (M_c(x, \eta_i) - \overline{M_c^p(x)})^2 \tag{5}$$

- One signal method to visualize the input patterns that activate neurons in higher layers is **Guided Backprop (GB) [20]**. GB achieves this by modifying the backpropagation process, which computes the gradient of the neuron activation with respect to the input features. Specifically, GB stops the flow of gradients at a ReLu gate when the gradient is less than zero.
- **Integrated Gradients (IG) [21]** is an example of an attribution method that assigns significance to input features by breaking down the output activation into contributions from each input feature.
- **X Gradients (XG)** is an unpublished XAI method that is still under development by the author of this project report. The technical detail will not be presented in this report.

C. Adversarial Attack and XAI

A comprehensive thesis [22] in 2019 bridges the gap between the XAI and the Adversarial Attacks. It reveals that the use of explainable AI metrics has been explored in training filter neural networks to identify adversarial examples. The goal was to determine whether these techniques could effectively identify patterns in adversarial examples that do not exist in the original data, thereby increasing the networks' robustness. Evidence has been provided that shows the linkage between the explainable techniques and the Adversarial Attack is a promising direction for searching the defense solutions. However, the thesis has emphasized the significant challenge of protecting neural networks from adversarial examples.

In this report, we try to establish the initial study of how adversarial samples generated from different attack techniques affect other gradient-based explanations' performance. We hope to reveal the clues of defense solutions by observing such performance variations and further utilize the findings to defend against attacks.

III. Methodology

To evaluate the performance loss of the explanation, we leverage the ground-truth segmentation for the object detection task as the ground-truth saliency map explanation. By default, the saliency value range from 0.0 to 1.0. The higher the value is, the more important the pixel is. The ground-truth segmentation projects only 0.0 and 1.0 saliency values where the object has the value 1.0. As Fig. 3 shows, the adversarial attack does affect the explanations generated by some XAI methods. We calculate the mean squared error (MSE) loss as the loss function between the ground-truth segmentation and the generated explanations. We further extend the loss to x metrics.

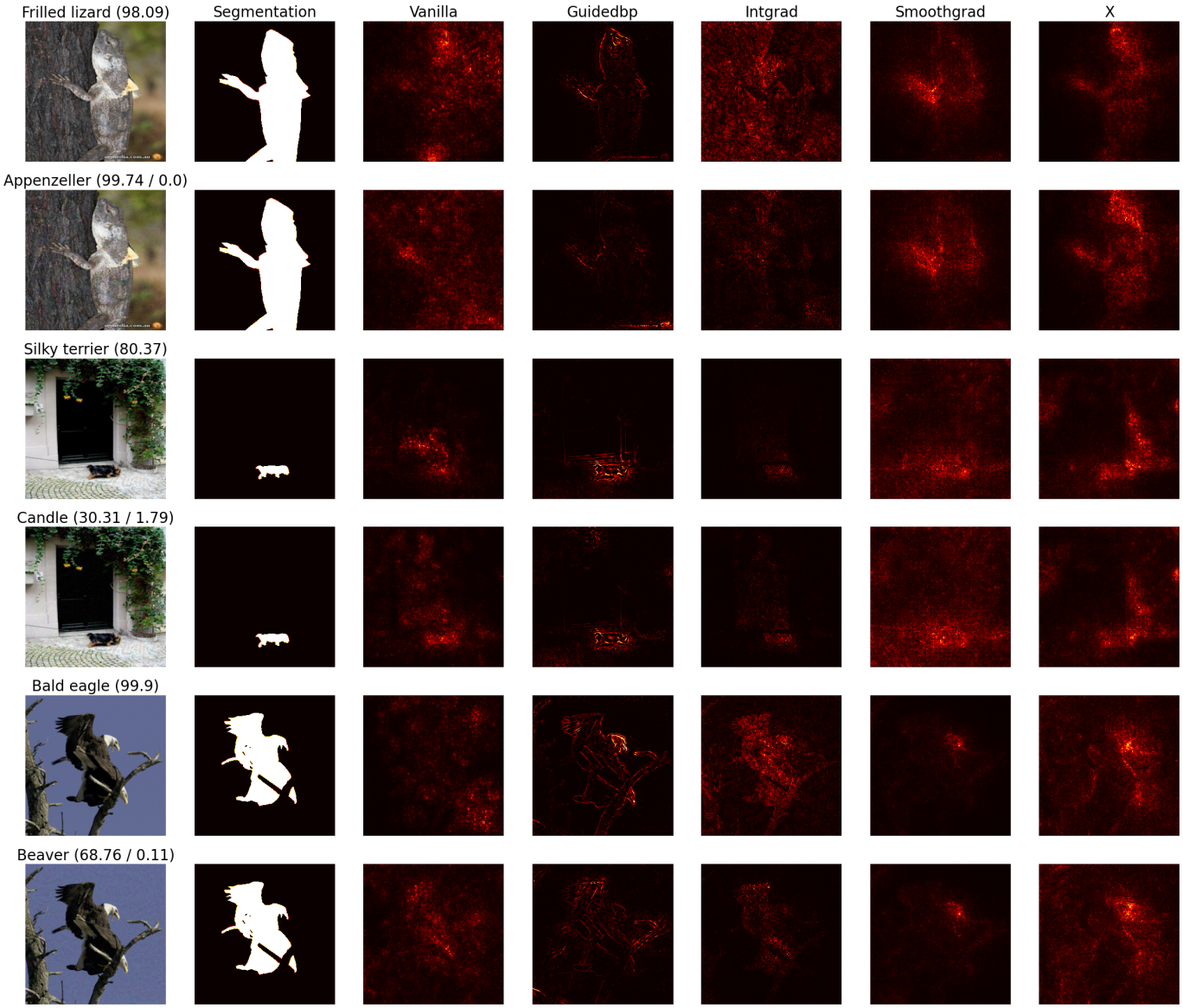

Fig. 3: Examples of the effects of the adversarial attack PGD. Three samples are shown here, with each having two rows. The first row explains the original image with its original prediction labeled on top of the original images: label (prediction confidence). The second row explains the adversarial sample with its prediction labeled on top of the adversarial images: new label (prediction confidence on the new/original labels). Some explanation focuses are shifted to the adversarial samples.

A. Loss of the Explanation (LOE)

We first evaluate the square root of the MSE loss with the 'sum' reduction of the explanation w.r.t another explanation as Equation. 6 shown:

$$LOE(x,y) = \sqrt{sum(\ell(x,y))}, \quad \ell(x,y) = L = \{l_1,\ldots,l_n\}^T, \quad l_n = (x_n - y_n)^2 \tag{6}$$

where x and y are from different explanations. This metric indicates how much the explanation differs from one another. We evaluate the XAI method's performance by calculating the LOE for the XAI and segmentation explanations. We evaluate how much an attack affects the XAI explanation by calculating

the LOE for the original and adversarial explanations.

B. Saliency Hit Rate(SHR)

We evaluate the hit rate of the saliency information of the explanation by calculating how much portion saliency value it projects to the ground-truth area of the segmentation explanation. Suppose the segmentation explanation S is a $m \times n$ matrix of 0.0 and 1.0, and the XAI explanation M is a $m \times n$ matrix of floating number range from 0.0 to 1.0. We filter the XAI explanation using the segmentation explanation as the mask as the following Equation. 7 is shown, and the resulting explanation is shown in Fig. 4.

$$\hat{M} = \begin{cases} M_{x,y} & \text{if } S_{x,y} = 1 \\ 0 & else \end{cases} \qquad x \in \{0, \ldots, m\}, y \in \{0, \ldots, n\} \tag{7}$$

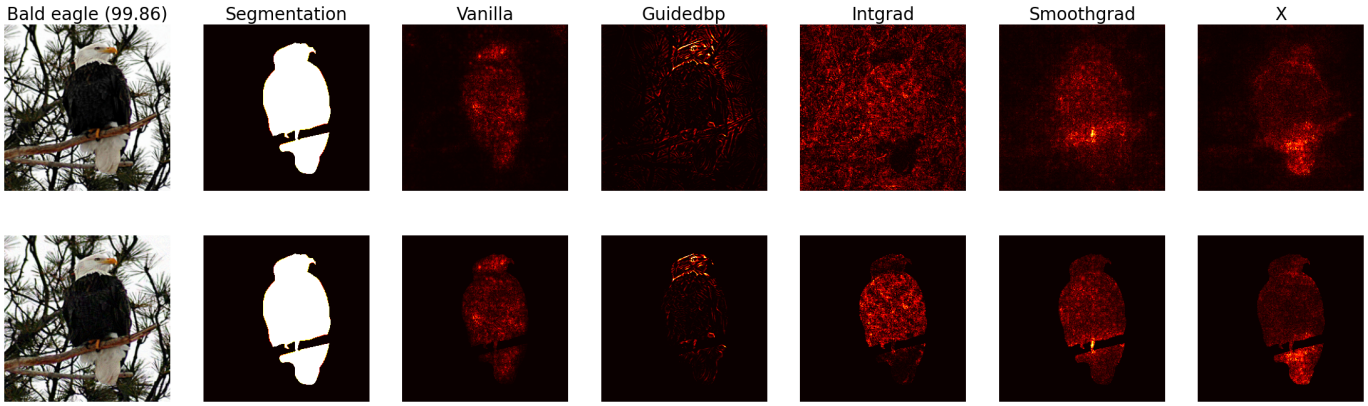

Fig. 4: Examples of the original explanation (first row) masked explanations (second row).

And then, we evaluate the Saliency Hit Rate (SHR) by summing up the saliency value of $\hat{M}$ and dividing it by the sum of the saliency value M.

$$SHR(M) = \frac{sum(\hat{M})}{sum(M)} \times 100 \tag{8}$$

This metric indicates how accurate an explanation focuses on the ground-truth object. The higher the value is, the more focused the explanation is. By what means focusing, we define it as (1) the heat (sum of the saliency value) on the ground-truth object region is higher; (2) the heat on the background region is lower. On the other hand, the more focused an explanation is, the less noisy it is visually.

C. Difference of the Saliency Hit Rate (DSHR)

With the evaluated Saliency Hit Rate(SHR), we assess the difference between the SHR of the original explanation and the adversarial explanation. We keep the sign since the positive DSHR indicates the adversarial explanation performs better than the original.

$$DSHR(M, M') = SHR(M) - SHR(M') \tag{9}$$

In conclusion, we define four metrics: (1) LOE of the original explanation w.r.t segmentation explanation. (2) LOE of the adversarial explanation w.r.t segmentation explanation. (3) LOE of the adversarial explanation w.r.t original explanation. (4) DSHR of the adversarial explanation w.r.t original explanation. The previous three metrics show how much the adversarial attack affects the XAI methods. The last metric

shows if the effect made by the attack is positive or negative. We accumulate enough data with the four defined metrics to perform statistical analysis through different experiment settings. We perform the XAI method on n numbers of images and take the mean value as the overall evaluation.

IV. Experiment

In this section, we first introduce the setup of the experiments. We then present the result and provide the analysis for each experiment. The four metrics organize the analysis.

A. Experiment Setup

We select the pre-trained ResNet-50 model published by the Pytorch team. Our target dataset is the ImageNet-1000 [23] dataset. Hence our target model is an image classification model. Moreover, we select the ImageNet-S [24] as the ground-truth segmentation explanation for our evaluation. For each experiment, we perform the XAI method on 1,000 samples. Next, we select the following eight attack techniques: (1) PGD, (2) TPGD, (3) EOTPGD, (4) FGSM, (5) FFGSM, (6) RFGSM, (7) BIM, and (8) Jitter. The attacks are implemented by [25]. For XAI method selection, we select and implement five methods, including: (1) Guided Backpropagation (GB), (2) Integrated Gradients (IG), (3) SmoothGrad (SG), (4) Vanilla Gradients, and our developed method (5) X Gradients (XG).

We set three more experiments for each attack with different attack function parameters, namely, perturbation rate as (1) 2/255, (2) 4/255, and (3) 8/255. This is to observe the effect of the percentage of the perturbation on the explanation.

B. Result and Analysis for the three LOE Metrics

Fig. 5a shows the LOE statistical results for the original explanation w.r.t the segmentation explanation. The result reveals the properties of each XAI method. Method GB and method Vanilla Gradients do not involve the introduction of randomness. Hence all of their original explanations are stable. On the other hand, methods IG, SG, and XG generate explanations with randomness. By observing the loss values, we find that method XG performs better than all the other methods with the loss values range in [100.344, 100,768]. This means method XG is closer to the ground-truth explanation. Other methods' loss values are over 110.0.

Fig. 5b shows the LOE statistical results for the adversarial explanation w.r.t the segmentation explanation. Compared with Fig. 5a, the result shows that method XG is least affected by all the examed attacks. Methods SG and Vanilla Gradients lose more than 1.0 similarity on average. Methods GB and IG lose more than 2.0 similarity on average.

Fig. 6a shows the LOE statistical results for the adversarial explanation w.r.t the original explanation. Focusing on the "Fold" column, the result shows that the more pixels are perturbed, the more difference between the original and original explanations. This observation is consistent through different XAI methods and attack techniques except for the IG method. Attack methods FFGSM and FGSM affect XAI methods SG and XG more than other XAI methods. Attack methods RFGSM and TPGD affect XAI methods GB and IG more than other XAI methods.

C. Result and Analysis for the DSHR Metric

As Fig. 6b is shown, XAI method GB is most affected by the attacks. In the last fold of experiments, the explanations of GB are contaminated with a loss of more than 12% of saliency hit rates. This means that method GB is vulnerable to the selected attacks, especially BIM, PGD, and EOTPGD attacks. By contrast, methods XG and SG are robust against the attacks with less than 1.0% loss on the saliency hit rates. This might be a direction for developing the robust XAI method or Deep Learning model against adversarial attacks. Further observing the IG method, it performs better on the adversarial samples than the original sample. This might be a direction of improving the IG method by leveraging the idea of some of the attacks.

Attack	Fold	GB	IG	SG	Vanilla	XG
BIM	1	115.793846	114.057278	110.866270	112.863811	100.491458
EOTPGD	1	115.793846	114.057278	110.562017	112.863811	100.491458
FFGSM	1	115.793846	113.955431	110.930944	112.863811	100.459081
FGSM	1	115.793846	114.014735	110.805230	112.863811	100.607649
Jitter	1	115.793846	114.158106	110.834661	112.863811	100.640205
PGD	1	115.793846	114.014735	110.805230	112.863811	100.607649
RFGSM	1	115.793846	114.014735	110.775635	112.863811	100.607649
TPGD	1	115.793846	114.026729	111.029567	112.863811	100.500079
BIM	2	115.793846	114.011647	110.840752	112.863811	100.378224
EOTPGD	2	115.793846	114.011647	111.036519	112.863811	100.378224
FFGSM	2	115.793846	114.287521	110.868092	112.863811	100.714094
FGSM	2	115.793846	114.180195	110.870107	112.863811	100.371946
Jitter	2	115.793846	114.189880	111.016036	112.863811	100.482890
PGD	2	115.793846	114.180195	110.870107	112.863811	100.371946
RFGSM	2	115.793846	114.180195	110.881980	112.863811	100.371946
TPGD	2	115.793846	114.254470	111.149206	112.863811	100.814930
BIM	3	115.793846	113.877703	110.873532	112.863811	100.603477
EOTPGD	3	115.793846	113.877703	111.162329	112.863811	100.603477
FFGSM	3	115.793846	113.975856	111.055328	112.863811	100.344235
FGSM	3	115.793846	114.160896	110.743168	112.863811	100.528882
Jitter	3	115.793846	114.112783	111.189110	112.863811	100.768226
PGD	3	115.793846	114.160896	110.743168	112.863811	100.528882
RFGSM	3	115.793846	114.160896	110.949185	112.863811	100.528882
TPGD	3	115.793846	114.142884	110.962009	112.863811	100.727274

(a) Original w.r.t the segmentation explanation.

Attack	Fold	GB	IG	SG	Vanilla	XG
BIM	1	116.922461	117.940184	110.985550	113.321831	100.347517
EOTPGD	1	116.999464	118.019530	111.073182	113.348119	100.410106
FFGSM	1	117.225626	116.658002	111.137169	113.266017	100.673925
FGSM	1	117.258758	116.542064	111.044927	113.565012	100.424736
Jitter	1	116.685538	117.551989	110.869822	113.275648	100.472792
PGD	1	116.922461	117.932084	111.104458	113.321831	100.570255
RFGSM	1	117.180938	118.064845	111.117185	113.286233	100.545298
TPGD	1	116.382719	116.529215	111.161692	112.989881	100.648020
BIM	2	117.604327	117.859139	111.068855	113.522439	100.527322
EOTPGD	2	117.332447	117.717099	111.106081	113.540726	100.741400
FFGSM	2	117.170567	116.399929	110.885496	112.045341	100.679932
FGSM	2	117.214800	116.285565	110.873047	112.657705	100.938006
Jitter	2	116.822864	117.189704	110.873830	112.706879	100.250888
PGD	2	117.604327	117.878094	110.960771	113.522439	100.625368
RFGSM	2	117.429163	117.933218	110.840899	113.528459	100.880644
TPGD	2	116.555951	116.327464	110.695563	113.198605	100.383329
BIM	3	117.772319	117.203639	111.019553	113.262198	100.863672
EOTPGD	3	117.749063	117.327691	110.932201	113.384573	100.798926
FFGSM	3	117.078589	116.279737	111.166459	112.142702	101.227824
FGSM	3	116.933863	116.301685	110.933141	111.668685	100.682220
Jitter	3	116.598965	116.570735	111.359231	112.438519	100.290241
PGD	3	117.772319	117.260810	111.157455	113.262198	100.633654
RFGSM	3	117.501032	117.443073	111.128948	113.138530	100.753144
TPGD	3	117.095322	116.188975	111.000266	113.175381	100.685589

(b) Adversarial w.r.t the segmentation explanation.

Fig. 5: Loss of the Explanation (LOE) statistical results

Attack	Fold	GB	IG	SG	Vanilla	XG
BIM	1	13.607543	18.208344	13.661258	15.320623	12.831740
EOTPGD	1	13.814624	18.138084	13.714072	15.810758	12.768977
FFGSM	1	13.704299	17.715373	13.307806	16.016333	12.822432
FGSM	1	13.866310	17.738358	13.557628	15.809415	12.741164
Jitter	1	12.544081	17.551088	13.620845	15.094084	12.847547
PGD	1	13.607543	18.789093	13.590300	15.320623	12.431985
RFGSM	1	14.135081	18.856355	13.444724	15.899250	12.605482
TPGD	1	14.400961	18.533515	13.246741	15.530870	13.009240
BIM	2	19.100644	18.656642	13.554666	16.881320	13.251301
EOTPGD	2	18.746420	18.754840	13.486316	16.982744	13.220590
FFGSM	2	16.734895	16.954504	13.633847	17.295656	13.666398
FGSM	2	16.801238	17.586090	13.745455	16.713335	13.321313
Jitter	2	16.211310	17.796045	13.437639	17.001359	13.154608
PGD	2	19.100644	18.389673	13.732462	16.881320	13.101901
RFGSM	2	18.900580	18.600337	13.761820	16.909235	13.238989
TPGD	2	21.100610	18.606214	13.463350	16.581541	13.502799
BIM	3	24.624886	18.324974	13.650643	17.804239	14.079373
EOTPGD	3	23.987310	18.329679	13.869226	17.724959	14.078884
FFGSM	3	19.632885	17.529861	13.599333	17.658245	14.721568
FGSM	3	20.027678	17.517051	14.234844	17.578739	14.976948
Jitter	3	19.707967	17.937442	13.466040	17.300817	13.541728
PGD	3	24.624886	18.306485	13.867807	17.804239	14.061097
RFGSM	3	24.349141	18.328599	13.764412	17.900521	13.999406
TPGD	3	28.508126	19.221437	13.509417	17.507890	14.052314

(a) LOE of original w.r.t the adversarial explanation.

Attack	Fold	GB	IG	SG	Vanilla	XG
BIM	1	6.790561	-4.115478	0.199247	1.664576	-0.152686
EOTPGD	1	7.137347	-3.806131	0.136846	2.354339	-0.079035
FFGSM	1	4.316643	-3.210795	0.008938	5.310044	0.133448
FGSM	1	4.478016	-2.806918	0.020006	5.448060	0.294501
Jitter	1	3.304935	-4.759280	-0.095476	3.173008	0.117629
PGD	1	6.790561	-4.171238	0.065420	1.664576	0.069811
RFGSM	1	7.774939	-3.135000	-0.155076	3.443888	0.166965
TPGD	1	6.419704	-1.967974	0.147707	2.058378	-0.100135
BIM	2	13.310742	-2.948358	0.214178	5.840425	0.123120
EOTPGD	2	12.507534	-2.975871	0.043512	5.420686	0.009708
FFGSM	2	5.556812	-4.461738	0.235946	4.449846	0.359268
FGSM	2	5.785442	-3.944069	0.245583	4.836813	0.369461
Jitter	2	6.207170	-4.482866	0.048874	4.031335	0.060499
PGD	2	13.310742	-2.653746	0.359659	5.840425	0.146997
RFGSM	2	12.695112	-2.380749	0.357084	5.222208	0.053466
TPGD	2	11.241770	-1.338186	0.448621	4.347000	0.171327
BIM	3	17.515766	-2.976041	0.406748	6.920039	0.364680
EOTPGD	3	17.028925	-2.906339	0.367609	6.683433	0.403366
FFGSM	3	6.938004	-4.638993	0.429371	4.286076	0.861303
FGSM	3	6.448547	-5.077780	0.576332	3.386305	0.895694
Jitter	3	6.540667	-5.590048	-0.045755	4.225366	-0.140524
PGD	3	17.515766	-3.188883	0.445658	6.920039	0.363324
RFGSM	3	16.855038	-2.884507	0.437638	6.681753	0.415227
TPGD	3	15.857811	-0.161458	0.241428	5.032040	0.241664

(b) DSHR result

Fig. 6: Loss of the Explanation (LOE) and the Difference of the Saliency Hit Rate (DSHR) statistical results

V. Conclusion

This report touches on the shallow of the joint topic of Adversarial Attacks and explainable AI in the Computer Vision domain. We define loss-based evaluation metrics for observing the performance loss when the XAI methods react on the adversarial samples. We perform experiments on eight attacks on five XAI methods and collect the statistical results of the defined metrics. The result shows the vulnerability and robustness of similar XAI methods against Adversarial Attacks. Hence, it leads to an optimistic view of the relationship between the attack and the explainability. The finding can further stem the study of the attack defense and the robustness of the XAI method against the attacks.

References

[1] J. Vadillo, R. Santana, and J. A. Lozano, "When and how to fool explainable models (and humans) with adversarial examples," *arXiv preprint arXiv:2107.01943*, 2021.

[2] X. Yuan, P. He, Q. Zhu, and X. Li, "Adversarial examples: Attacks and defenses for deep learning," 2018.

[3] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[4] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," 2017.

[5] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," 2015.

[6] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," 2017.

[7] N. Akhtar, A. Mian, N. Kardan, and M. Shah, "Advances in adversarial attacks and defenses in computer vision: A survey," 2021.

[8] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 2019.

[9] G. Sriramanan, S. Addepalli, A. Baburaj, and R. V. Babu, "Guided adversarial attack for evaluating and enhancing adversarial defenses," 2020.

[10] M. Cheng, Q. Lei, P.-Y. Chen, I. Dhillon, and C.-J. Hsieh, "Cat: Customized adversarial training for improved robustness," 2020.

[11] C. Xie, M. Tan, B. Gong, J. Wang, A. Yuille, and Q. V. Le, "Adversarial examples improve image recognition," 2020.

[12] G. F. Elsayed, I. Goodfellow, and J. Sohl-Dickstein, "Adversarial reprogramming of neural networks," 2018.

[13] K. Sakaguchi, R. L. Bras, C. Bhagavatula, and Y. Choi, "Winogrande: An adversarial winograd schema challenge at scale," 2019.

[14] M. A. Jalwana, N. Akhtar, M. Bennamoun, and A. Mian, "Attack to explain deep representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9543–9552.

[15] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[16] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," 2016.

[17] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[18] D. Smilkov, N. Thorat, B. Kim, F. Viégas, and M. Wattenberg, "Smoothgrad: removing noise by adding noise," 2017.

[19] L. Richter, A. Boustati, N. Nüsken, F. J. R. Ruiz, and Ömer Deniz Akyildiz, "Vargrad: A low-variance gradient estimator for variational inference," 2020.

[20] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.

[21] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *International conference on machine learning*. PMLR, 2017, pp. 3319–3328.

[22] H. Stiff, "Explainable ai as a defence mechanism for adversarial examples," 2019.

[23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 248–255. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/5206848/

[24] S. Gao, Z.-Y. Li, M.-H. Yang, M.-M. Cheng, J. Han, and P. Torr, "Large-scale unsupervised semantic segmentation," 2022.

[25] H. Kim, "Torchattacks: A pytorch repository for adversarial attacks," *arXiv preprint arXiv:2010.01950*, 2020.