Performance evaluation of Explainable AI methods against adversarial noise

L. M. Fenoy, A. Ciontos

***Abstract*—Recent work in machine learning has yielded in algorithms with high performance and accuracy. However, in critical areas such as medicine, finance or law, these algorithms are not yet fully trusted. The reason for this is their "black-box" nature. Meaning, when they fail, there is no clear reason for the failure. To overcome this issue, explainable AI (XAI) algorithms have been developed to add an extra layer of explainability towards AI. But with adversarial attacks at hand, even these algorithms become vulnerable. The aim of this paper is to study the effect of Fast Gradient Sign Method (FGSM) adversarial attack on two recent XAI algorithms, namely Similarity Difference and Uniqueness (SIDU) and Gradient-weighted Class Activation Mapping (Grad-CAM). Furthermore, by employing an eye tracker, we analyse how human eye fixation on natural images can be perceived and compared to the XAI saliency map. Our findings are that even though initially GradCam performs better than SIDU, when compared to the fixation maps as a ground truth, when it comes to noise, the results switch, thus SIDU is in fact more robust to adversarial attacks.**

***Index Terms*—XAI, Adversarial attacks, FGSM, Natural Images**

I. Introduction

The rise of machine learning has greatly impacted our society. Not only in scientific or technological areas, but also in medical, financial and even entertainment applications. Potentially, any application which involves analysing big data, is a suitable candidate for machine learning algorithms to take over.

However, while these algorithms can find trends and patterns in the data with great accuracy and repeatability, it is hard to understand all the underlying processes that lead to a specific decision. Resulting in the term "black box" being frequently used to denominate the inner workings of machine learning algorithms. Due to this "black box" characteristic, when a machine gives the wrong prediction, we are usually at a loss determining whether it was caused due to a bias in the data, a fault in the model's architecture, or even deliberate attacks designed to alter the model prediction.

Not being able to properly understand why machines make the decisions they do, creates a level of distrust which becomes inherently more so, when a model's prediction has a direct impact on human lives, for instance in areas such as medicine, finance or law. [1] [2]

While there are ethical implications involved with whether or not a machine should be responsible for making such decisions, from a technical point of view, AI algorithms provide enough benefits to motivate their use. In order to establish trust between humans and AI algorithms, a new branch of artificial intelligence has recently started to gain traction. Namely Explainable AI (XAI) which focuses on interpretability assessment criteria (such as reliability, causality and usability), often by generating visualisations comprehensible for a human. [3]

Nonetheless, while these algorithms have proven to generate legible explanations on different datasets, recent studies have concluded that adversarial attacks render most explanations generated by XAI methods obsolete. [4] [5] [6] Adversarial attacks introduce perturbations in the form of subtle modifications to the data to be analysed, which cause the models to make inaccurate predictions. These attacks are especially dangerous when they are undetectable by the human eye. Therefore, to maintain trustworthiness, XAI algorithms should be robust to such attacks.

A. Contribution

The dataset used for the present research consists of natural images. All images of objects naturally found in our surroundings will be referred to as natural images. In this regard, the contributions presented in this paper will be the following:

- Analysis of natural images. This part will analyse human attention in terms of fixation maps, when presented with natural images. This is especially important, because the aim of XAI algorithms is to produce comprehensive explanations, they should resemble how humans understand these images.
- Analysis of XAI explanation heatmaps compared to fixation maps. This part will analyse the similarity between the way humans perceive images compared to how XAI methods explain them.
- Analysis of the effect of adversarial noise on XAI methods. This part will compare how two different XAI algorithms react to different levels of adversarial noise and whether or not their generated explanation remain consistent with the original fixation maps.

II. Related work

Since the term XAI was coined by DARPA [7] as an initiative to unravel the black box characteristic behind machine learning, there have been multiple interpretations on how models should enhance their interpretability. Some applications such as LIME [8], provide local explanations in the form of linear approximations. This technique aims to highlight which features had the most impact on the model decision making. Other applications such as Activation Atlas

[9], aim to understand how networks ”see” images at different layers, they do so by providing feature visualisations of averaged activations. Whereas other XAI approaches, such as Grad-CAM [4], produce visual explanations.

These XAI methods have contributed tremendously to establishing trust between humans and machines. However, they remain vulnerable to adversarial attacks. Thus, since the algorithm can be disturbed by small perturbations, the explanations become unreliable. [5] [6] There are two main types of adversarial attacks, white-box attacks black-box attacks. The former one, unlike the latter one, requires access to model parameters. Another way to classify adversarial attacks is whether they are targeted or non-targeted. Targeted attacks aim for misclassification to a specifically defined class, while a non-targeted attack forces the algorithm to misclassify the input.

Examples of white-box attacks include among others, methods such as the Fast Gradient Sign Method (FGSM) [10], or Projected Gradient Descent (PGD) [11]. The Fast Gradient Sign Method works by introducing a small amount of noise to the image, which is indistinguishable by the human eye. The direction of this noise is the same as the gradient of the cost with respect to the input data. Similarly, the Projected Gradient Descent (PGD) works by iteratively applying FGSM to the image, thus generating an adversarial example, which is then repeatedly projected as a valid example. Furthermore, examples of black-box attacks, which are unrelated to the model parameters are Carlini-Wagner attack (CW) [12], or DeepFool. [13] Most of the aforementioned methods perform pixel-wise operations on images, meaning all pixels are changed slightly. However, there are methods that introduce perturbations only in a specific location of the image. An example of such attack is the adversarial patch. [14]

III. Analysis of natural images

In order to analyse what areas in a natural image attract most attention for a human observer, we set up an eye tracking experiment. An eye tracker is a device that points near infrared light into the pupil. The light is then reflected inside the optical system, which results in detectable reflections between the cornea (outer-most layer of the eye) and the pupil. This reflection is then recorded by an image sensor. This measurement yields the point of gaze, in other words, the direction the person is looking towards. There are different metrics to gather appropriate eye tracking data, out of which we use fixations for establishing areas of interest. Fixations are clusters of gaze points. [15] The gaze points are collected according to the frequency of sampling, i.e. 60 Hz. Hence, fixations are used to measure the distribution of visual attention. In order to visualise the visual attention, we generate fixation maps.

This method is chosen because of its similarity to XAI visualisations, which often come in forms of heatmaps representing salient areas in an image, therefore becoming a relevant metric to compare the similarity between the way humans perceive images compared to how the XAI algorithms explain them.

IV. Fast gradient sign method (FGSM)

As presented in Section II, there are multiple methods which can be used to generate adversarial attacks. However, when choosing a specific method, an important fact to consider is whether or not the network’s architecture and parameters are known. Currently, most successful attacks are white box attacks, specifically, gradient based attacks. Some examples of such attacks are FGSM and PGD. PGD is an iterative application of FGSM, it is stronger than FGSM, but the process is also more complex and time consuming. Therefore, because of its simplicity and effectiveness at the same time, we have chosen the Fast Gradient Sign Method (FGSM).

FGSM works by adding carefully calculated noise to an image. The direction of this noise is the same as the gradient of the cost with respect to the input data. This is given by the direction of the gradient, in other words, the Gradient Sign (+/-). The amount of noise can be controlled by a coefficient, epsilon. When this coefficient is applied correctly, it will alter the model prediction while still being undetectable to the human eye. The following formula stands at the base of generating FGSM noise:

$$adv\_x = x + \varepsilon sign(\quad _x J(\ ; x; y)) \tag{1}$$

Where: adv_x = Adversarial image, x = Original image, ε = Coefficient, $\quad _x$ = Gradient, J = Loss, $\quad$ = Model parameters, y = Input label.

In order to better demonstrate both the robustness and explainability of different XAI algorithms, we experimentally defined three optimal noise coefficients (ε). The chosen values are $\varepsilon = 0.007; \varepsilon = 0.05$ and $\varepsilon = 0.1$. The first value is optimal because it is small enough to pass unnoticeable by the human eye, but large enough to affect the algorithms, as also demonstrated by [10]. The other two coefficients are chosen for experimental purposes. Figure 1 showcases a natural image with FGSM type noise added on the three chosen values and the resulted predictions.

(a) Original: Flamingo (b) $\varepsilon = 0.007$: American egret (c) $\varepsilon = 0.05$: Nematode (d) $\varepsilon = 0.1$: Nematode

Fig. 1: Example of a natural image in its original form and also with three different levels of noise, together with the corresponding predictions

V. XAI Methods & Hypothesis

This paper compares two XAI algorithms. The first one being Gradient-weighted Class Activation Mapping (Grad-CAM) [4]. Grad-CAM is a method which generates visual explanations via gradient based localization. To do so, it extracts the gradients from the last convolution layer of the network. The intuition behind this method is that the layer

prior to the classification retains the information of feature relevance while maintaining spatial relations, and therefore it can generate a heatmap (based on a weighted combination of activation maps dependent on gradient score) which highlights the features with a positive influence for the specific class which is chosen as the prediction.

The second XAI method evaluated is Similarity Difference and Uniqueness Method (SIDU) [16]. This method generates a heatmap based on two values: Similarity difference and Uniqueness. First, a heatmap of the most salient areas of an image is generated by calculating the similarity difference between sets of feature activation maps. Secondly, it evaluates feature map uniqueness. This step calculates how different a specific feature map is from the others. If a feature map is unique, then it will be labelled as more salient and have a higher weight. The final score that gives the feature importance is given by the dot product between the two values, which is then used to calculate the weighted sum of all feature activation image masks. and generate the visual explanation.

Both methods generate their heatmaps using information from the last convolution layer of the network and produce visual explanations which highlight the pixels that contribute to the class prediction. However, the way these heatmaps are generated greatly differs from each other. Our hypothesis is that since Grad-CAM depends on gradient values to generate its visual explanations, the FSGM will have a great impact on the heatmaps. On the other hand, since SIDU is a gradient free method, we believe that it will be robust to this kind of adversarial attack and be able to generate unaltered visual explanations.

VI. Experiments and results

A series of experiments have been carried out to compare robustness of XAI methods.

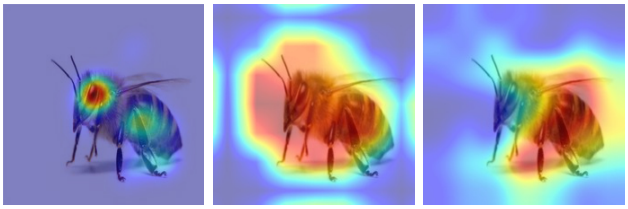
A. The dataset

The dataset used for the following experiments consists of 100 natural images, distributed as 10 images belonging to 10 different classes defined in ImageNet. [17] The images in the dataset are all RGB and resized to 224x224 pixels. This dataset is used both in the eye tracking experiments as well as the XAI algorithms.

B. How do humans and machines perceive natural images

As previously mentioned, in order for XAI explanations to be comprehensible to humans, they need to resemble how humans perceive images. Therefore, we need to collect data on humans' understanding of natural images. This is done using an eye tracker. The eye tracker records aggregated eye fixations across multiple participants. These eye fixations are then represented as the most salient areas in the fixation map.

For experimenting with eye tracker measurements, 5 subjects have been tested. Each image from the dataset was displayed in arbitrary order for 3 seconds as demonstrated by [18]. Finally, an aggregated heatmap of fixations gathered from all participants has been generated for each image. Figure 2 presents the results from the eye tracker study (human perception) as well as machine perception counterparts generated by SIDU and Grad-CAM.



(a) Eye tracker (b) SIDU (c) Grad-CAM

Fig. 2: Most salient areas in the same image, captured with different methods

In [19], Das et.al. concluded that machine learning algorithms do not define salient areas in an image the same way that humans do. This can be appreciated in Figure 2, where the results from the eye tacker show concentrated fixation points, whereas both XAI generated heatmaps show large areas of interest.

However, in order to quantify the resemblance between these heatmaps, we use the Kullback–Leibler divergence (KL divergence). [20] This method is used as a metric for estimating overall dissimilarity between two distributions. To compare the heatmaps from the eye tracker and each XAI method individually, we need to measure the dissimilarity between the saliency maps' probability distribution (SalMap) from SIDU and Grad-CAM and the human eye fixation probability distribution (Emap) as described by Eq.2. [21]

$$KLDiv = \sum_{x=1}^{X} E_{map}(x) \; log(\frac{E_{map}(x)}{SalMap(x) + } +) \quad (2)$$

Where X is the number of pixels and is a small coefficient to avoid log and division by zero. The KL Divergence will return scores between 0 and 1, the closer the score to 0, the more similar the distributions. We then use another metric to asses whether or not consistent patterns emerge in the heatmaps comparisons. For this purpose, we use Spearman's correlation, as presented by [19]. This is a non-parametric measure that analyses whether or not the relationship between two datasets is monotonic. This metric varies between -1 and 1, where a score of 0, represents no correlation. The sign shows whether the datasets are positively or negatively related, and the values represent how strong the relation is as mapped in the list bellow [22]:

- Very weak: 0.00 to 0.19
- Weak: 0.20 to 0.39
- Moderate: 0.40 to 0.59
- Strong: 0.60 to 0.79
- Very strong: 0.80 to 1.00

The expression bellow calculates this correlation, where u_i and v_i are ranks of the values collected from the two data sets, in this case between eye tracker measurements and XAI methods.

$$rs = \frac{n\sum_{i=1} u_i v_i - (\sum_{i=1} u_i)(\sum_{i=1} v_i)}{\sqrt{[n\sum_{i=1} u_i^2 - (\sum_{i=1} u_i)^2][n\sum_{i=1} v_i^2 - (\sum_{i=1} v_i)^2]}} \quad (3)$$

The results show that the KL divergence between the eye tracker measurements and SIDU has a value of 0.90, while between the eye tracker measurements and Grad-CAM it has a value of 0.81. For the same scenarios, Spearman's correlation returns 0.18 and 0.20 respectively. Therefore we can conclude that, on average, Grad-CAM's explanations are closer to the ground truth when compared to SIDU. The results are presented in Table I for the KL divergence and Table II for Spearman's correlation.

	SIDU + FGSM				Grad CAM + FGSM			
Noise Levels	**0**	**0.007**	**0.05**	**0.1**	**0**	**0.007**	**0.05**	**0.1**
Eye Tracker	0.90	0.87	0.75	0.71	0.81	1.54	1.37	1.63
SIDU	-	0.56	0.40	0.37	-	-	-	-
Grad CAM	-	-	-	-	-	1.17	0.99	1.17

TABLE I: Averaged KL divergence of SIDU and Grad CAM for different noise levels

	SIDU + FGSM				Grad CAM + FGSM			
Noise Levels	**0**	**0.007**	**0.05**	**0.1**	**0**	**0.007**	**0.05**	**0.1**
Eye Tracker	0.18	0.26	0.22	0.14	0.20	-0.11	-0.17	-0.16
SIDU	-	0.33	0.35	0.29	-	-	-	-
Grad CAM	-	-	-	-	-	-0.04	-0.10	-0.04

TABLE II: Averaged Spearman's correlation of SIDU and Grad CAM for different noise levels

C. How do XAI saliency maps deviate from fixation maps after applying FGSM generated noise?

For this experiment, FGSM noise with different epsilon levels as defined in Section IV has been added to the dataset. Then, we evaluated both SIDU and Grad-CAM, using the contaminated data. In this test, we want to analyse how robust the XAI methods are against an adversarial attack in terms of generating reliable explanations. Reliable explanation are defined in terms of similarity to the fixation maps. To collect the results we calculate the average KL divergence and Spearman's correlation for all pairs of maps as follows:

$$D = (\sum_{i=1} KLDiv(E_{map}(i); SalMap_{adv\_x}(i)))/I \quad (4)$$

$$S = (\sum_{i=1} rs(E_{map}(i); SalMap_{adv\_x}(i))/I \quad (5)$$

Where I is the total number of images and $SalMap_{adv\_x}$ is the saliency map obtained by the explanation of the adversarial examples. This operation is then performed for all noise levels. $KLDiv$ and rs are defined in Equation 2 and 3 respectively.

From Table I, it can be observed that when comparing the explanations from the noisy images to the eye tracker ground truth, the averaged KL-divergence (D) results from SIDU outperforms those of Grad-CAM by approximately a factor of 2. Regarding the averaged Spearman's correlation (S), the results from Table II show that the scores from Grad-CAM present a very weak negative correlation with the eye tracker heatmaps, whereas SIDU has a weak positive correlation. These results can also be visualised in Figure 10. Overall, SIDU seems to be more robust to adversarial noise compared to Grad-CAM.

D. How do saliency maps from adversarial examples deviate from original saliency maps?

As previously mentioned, human attention maps and machine heatmaps tend to differ. Therefore, we chose to compare how the XAI explanations behave before and after applying FGSM. Therefore we compare results from both SIDU and Grad-CAM algorithms with = 0 (no noise) as the ground truth with each subsequent noise level. The results should further prove the algorithm's robustness when faced with adversarial noise. The process is the same as described in Equations 4 and 5, except, the ground truth is switched from the fixation maps (E_{map}) to original saliency maps ($SalMap$).

When comparing the averaged KL-divergence results from Grad-Cam after the adversarial attacks to the original Grad-Cam heatmap, and the SIDU noisy heatmaps to original SIDU, once again, SIDU results are better by approximately a factor of 2, as it can be observed in Table I. Lastly, in Table II th averaged Spearman's coefficient shows a very weak negative correlation between FGSM perturbed images and a weak positive correlation for the SIDU images. Overall, SIDU still continues to be more robust to adversarial noise compared to Grad-CAM.

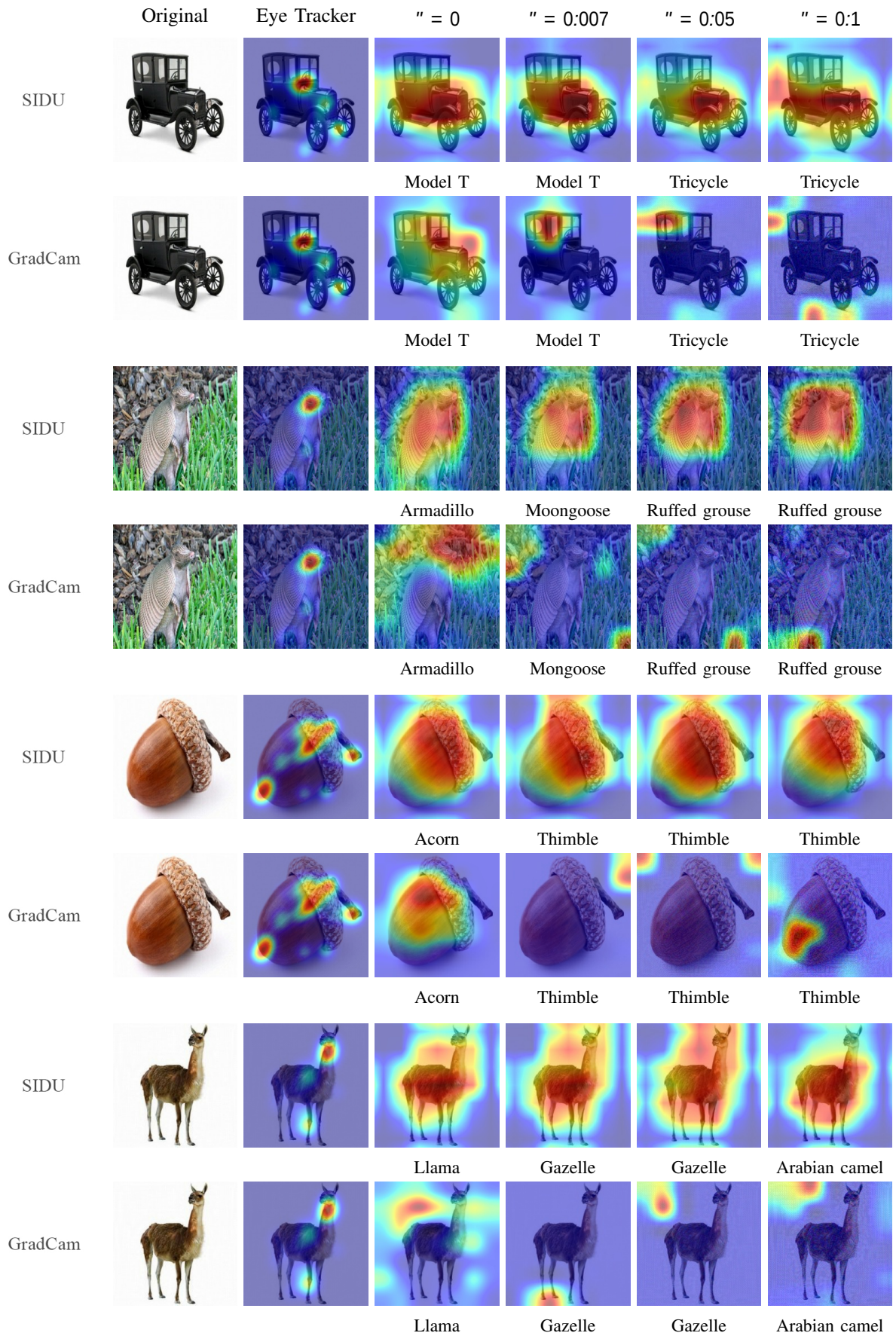


Fig. 10: Visual outputs and predictions for XAI methods with different noise levels

VII. Conclusion

In this paper, we carried out an analysis of natural images, to understand the similarity between human fixation maps gathered with an eye tracker, compared to XAI (Explainable Artificial Intelligence) saliency maps. Furthermore, we created adversarial examples using the FGSM algorithm, and used them as input for SIDU (Similarity Difference and Uniqueness) and Grad-CAM (Gradient-weighted Class Activation Mapping) XAI algorithms. The aim was to test how the explanations deviate from both fixation maps and original saliency maps prior to the adversarial attack. When generating explanations, both algorithms predict the same class and that is due to the fact that they have been implemented on the same model. Our results show that Grad-CAM visual explanations are more similar to the human fixation maps than SIDU explanations are. However, when FGSM noise is introduced, SIDU is more robust than Grad-CAM. We suspect this is true because SIDU, unlike Grad-CAM, is not a gradient based method.

References

[1] M. Dzindolet, S. Peterson, R. Pomranky, L. Pierce, and H. Beck, "The role of trust in automation reliance," *International Journal of Human-Computer Studies*, vol. 58, pp. 697–718, 06 2003.

[2] N. K. Eun-Jae Lee, Yong-Hwan Kim and D.-W. Kang, "Deep into the brain artificial intelligence in stroke imaging," *Journal of Stroke*, vol. 19, p. 277–285, 09 2017.

[3] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): Towards medical xai," 10 2019.

[4] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization," *CoRR*, vol. abs/1610.02391, 2016. [Online]. Available: http://arxiv.org/abs/1610.02391

[5] I. S. J. B. D. E. I. G. R. F. Christian Szegedy, Wojciech Zaremba, "Intriguing properties of neural networks," 12 2013.

[6] I. Evtimov, K. Eykholt, E. Fernandes, T. Kohno, B. Li, A. Prakash, A. Rahmati, and D. Song, "Robust physical-world attacks on machine learning models," *CoRR*, vol. abs/1707.08945, 2017. [Online]. Available: http://arxiv.org/abs/1707.08945

[7] D. M. Turek, "Explainable artificial intelligence (xai)," https://www.darpa.mil/program/explainable-artificial-intelligence, 2014.

[8] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should I trust you?": Explaining the predictions of any classifier," *CoRR*, vol. abs/1602.04938, 2016. [Online]. Available: http://arxiv.org/abs/1602.04938

[9] S. Carter, Z. Armstrong, L. Schubert, I. Johnson, and C. Olah, "Activation atlas," *Distill*, vol. 4, no. 3, Mar. 2019. [Online]. Available: https://doi.org/10.23915/distill.00015

[10] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 01 2015, pp. 1–10.

[11] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 06 2017.

[12] N. Carlini and D. A. Wagner, "Towards evaluating the robustness of neural networks," *CoRR*, vol. abs/1608.04644, 2016. [Online]. Available: http://arxiv.org/abs/1608.04644

[13] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," *CVPR*, 11 2016.

[14] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, "Adversarial patch," *CoRR*, vol. abs/1712.09665, 2017. [Online]. Available: http://arxiv.org/abs/1712.09665

[15] I. Mitsugami, N. Ukita, and M. Kidode, "Robot navigation by eye pointing," *Lecture Notes in Computer Science*, vol. 3711, pp. 256–267, 01 2005.

[16] T. B. M. Satya M. Muddamsetty, Mohammad N. S. Jahromi, "Sidu: Similarity difference and uniqueness method for explainable ai," accepted in IEEE International Conference on Image Processing(ICIP) - Jan 2020.

[17] P. U. Stanford Vision Lab, Stanford University. Imagenet. [Online]. Available: http://www.image-net.org/

[18] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *2009 IEEE 12th International Conference on Computer Vision*, 2009, pp. 2106–2113.

[19] A. Das, H. Agrawal, C. L. Zitnick, D. Parikh, and D. Batra, "Human Attention in Visual Question Answering: Do Humans and Deep Networks Look at the Same Regions?" in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.

[20] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *arXiv preprint arXiv:1604.03605*, 2016.

[21] N. Riche, M. Duvinage, M. Mancas, B. Gosselin, and T. Dutoit, "Saliency and human fixations: State-of-the-art and study of comparison metrics," 12 2013.

[22] O. Santiago, C. Carlos Alberto, A. Silva Neto, and J. Verdegay, *Computational Intelligence in Emerging Technologies for Engineering Applications*, 01 2020.