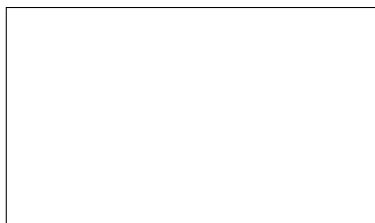# Graphical Abstract

Signed Saliency Maps

Oscar Llorente, Jaime Boal, Eugenio F. Sánchez-Úbeda

Highlights

Signed Saliency Maps

Oscar Llorente, Jaime Boal, Eugenio F. Sánchez-Úbeda

- Research highlight 1

- Research highlight 2

Signed Saliency Maps

Oscar Llorente[a], Jaime Boal[b], Eugenio F. Sánchez-Úbeda[b]

[a]*BMAS SA NDO SW R&D Unit B, Ericsson, Retama Ed 1 Torre Suecia, Madrid, 28045, Madrid, Spain*
[b]*Institute for Research in Technology (IIT), ICAI School of Engineering, Comillas Pontifical University, Santa Cruz de Marcenado, 26, Madrid, 28015, Madrid, Spain*

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Keywords: keyword one, keyword two
PACS: 0000, 1111
2000 MSC: 0000, 1111

1. Introduction

Since the introduction of AlexNet [1] in the ImageNet [2] competition there has been a revolution in the Computer Vision field, where almost in any problem deep learning architectures are the considered the state of the art. However, this type of technology has a big drawback, specially if it is used in sensitive domains, as medicine or autonomous vehicles, its explainability. Due to the high number of parameters (billions) and the complexity of these techniques, it is difficult to explain how a neural network make a certain prediction. To solve this problem, there are two main lines of research:

- Building more interpretable models.

- Developing techniques to explain the state of the art techniques.

There have been attempts to construct these interpretable models for the sensitive domains as in [3]. However, there are also many other techniques used for Computer Vision problems that offer great results and are not easy to explain, as the classical Convolutional Neural Networks (CNNs) [4]. To explaining these other models several techniques have been developed during the last years. One of the most used types are Saliency Maps.

Saliency Maps were introduced first in [5] to explain the classification of a neural network based on its derivatives. This type of methods are commonly tested in a multi-class classification problem that can be defined as the following:

$$class(I) = argmax_{c \in C} S_c(I), \tag{1}$$

being I the input image, C the set of all possible classes and S_c the classification score. Then, the original Saliency Map method is expressed as:

$$M_c(I) = max_{channel}\left\{\left|\frac{\partial S_c}{\partial I}\right|\right\}. \tag{2}$$

According to [5], the brightest points in a Saliency Map, i.e., the derivatives with a higher absolute value, are the points that, with a minimum change, affect the class score the most. Then, to derive a single value per pixel, the maximum function over the three pixels is computed. Even though this decision can seem arbitrary, this is the convention that has been followed in every paper about Saliency Maps.

However, the original Saliency Maps are very noisy. Therefore, in the following years many variations were suggested to try to improve this visualization and make this noise disappear, as in [6], [7] or [8]. By this, sharper visualizations were achieved, being these ones more similar to the object in the image and reducing the noise. Unfortunately, in [9] it was proved that some of these techniques were not showing information about what the neural network has learned, but features from the image, as an edge detector.

Nevertheless, there is no need of these techniques that improve the visualizations by projecting the image information: more knowledge can be unlocked from the gradients of the neural network without further modifications. First, as far as we know, the sign of the gradients has not been studied in any paper. This sign can have a meaning and a significant impact in the neural network behavior that researchers are trying to understand. Moreover, the type of problem being studied is a multi-class classification

2

problem, and a visualization that tries to explain that, should take into account all these classes and not only the one that is being predicted. In other words, the magnitude of the gradients can have different meaning if it is greater or smaller than the gradients for other classes.

2. The Sign in Saliency Maps

As pointed out in the last section, there is not any research about the meaning or impact of the sign in Saliency Maps. The only article that discuss briefly this aspect is [10]. In this paper it was explained that the raw value of gradients (without absolute value) is commonly used for MNIST dataset [11] but not in other datasets as ImageNet. The reason for this is that in the former it produces clearer pictures and in the latter is the opposite.

However, by taking the absolute value of all the pixels in the Saliency Map, information can be lost. It is not only valuable knowing which pixels are important but also their meaning. For example, it would be significant, for the explainability field, having a technique that can indicate which zones of the image should be brighter to improve the classification and which ones should be darker. This is in fact the information that the sign can provide. Moreover, if both information are combined in the same image without any distinctions between them, they can produce worse explanations. It can seem as if the model is not able to distinguish between zones that should be brighter and others that should be brighter, as for example an object in front of a darker background. Based on this, two sets of pixels can be distinguished in the image:

- Pixels that by increasing their value the classification of the model improves[1] since they have a positive value in the Saliency Map (positive gradient).

- Pixels that by decreasing their value the classification of the model improves since they have a negative value in the Saliency Map (negative gradient).

Regarding the reason for not using the raw values exposed in [10], it can be due to a high number of 0 values in the Saliency Map, since when

---

[1]It is important to note that improving the model classification means that the logit of the predicted class increases.

normalized with positives and negatives these pixels will be shining at a medium intensity. To solve this, in this paper the visualization in different images is proposed, having the following new techniques:

- Positive Saliency Maps:

$$M_c(I) = max_{channel}\Big\{ReLU\Big(\frac{\partial S_c}{\partial I}\Big)\Big\}. \tag{3}$$

- Negative Saliency Maps:

$$M_c(I) = max_{channel}\Big\{ReLU\Big(-\frac{\partial S_c}{\partial I}\Big)\Big\} \tag{4}$$

By separating them before taking the absolute value the information about what the pixels mean is not lost.

3. Multi-class Saliency Maps

As it was mentioned in Section 1, the problem being studied is a multi-class classification problem. However, currently all Saliency Map techniques used the predicted class to compute the derivatives. That does not take into account the information about other classes that the model is predicting. In other words, there can be another gradient for other class with a higher value than for the predicted class and the current techniques cannot show that to the user. The reason for this is that, if the absolute value of all gradients is taken, there cannot be any comparison with the gradients of other classes, since there is no information anymore that can indicate what means that a gradient is higher or lower than another.

However, going further in the analysis started in the last section about the sign allows to compare the gradients of the predicted class and the other ones. In fact, it improves the previous definitions about the different sets of pixels identified in the last section:

- Pixels that by increasing their value the classification of the model improves[2] are not always the positive ones. Since now other classes are

---

[2]Since now the other classes are taken into account, the definition of improving the classification is that the difference of the logit of the predicted class and the other logits is greater.

taken into account, only the pixels that have a higher gradient for the predicted class that for any other class, will improve the classification if their value is increased.

- Analogously, the pixels that by decreasing their value improve the classification are only the ones that have a lower gradient for the predicted class than for any other class.

Therefore, by comparing the Saliency Maps for the different classes the information about the two sets mentioned before can be extracted. Two new techniques are proposed to represent this:

- Active Saliency Map (increasing them improves the classification):

$$\hat{M}_c(I) = max_{channel}\Big\{M_c \prod_{q \in (C-c)} M_c > \frac{\partial S_q}{\partial I}\Big\}, \qquad (5)$$

being M_c the Saliency Map (original technique) for the predicted class c.

- Inactive Saliency Map (decreasing them improves the classification):

$$\hat{M}_c(I) = max_{channel}\Big\{M_c \prod_{q \in (C-c)} M_c < \frac{\partial S_q}{\partial I}\Big\}. \qquad (6)$$

4. Experiments

In this section the new proposed techniques will be evaluated qualitatively and quantitatively. With this purpose, a certain methodology was followed. First, two datasets are used to study the new methods, Cifar10 [12] and Imagenette [13]. The first one is a commonly used dataset for image classification and also for explainability. The second one is a subset of 10 classes of Imagenet. It is used this subset instead of the original Imagenet due to computing restrictions of hardware. Then, two models are trained for each dataset. The first is a CNN model based in a max-pooling several layers for CNNs, and average-pooling and a final linear layer. The second one is the classical Resnet18 [14]. They were all trained with a learning rate of 0.001 and the AdamW optimizer for 50 epochs. The results of the test set are the following:

	Cifar10	Imagenette
Basic CNN model	0.6307	0.6205
Resnet18	0.7569	0.8357

Table 1: Training results

(a) image   (b) original   (c) positive   (d) negative   (e) active   (f) inactive

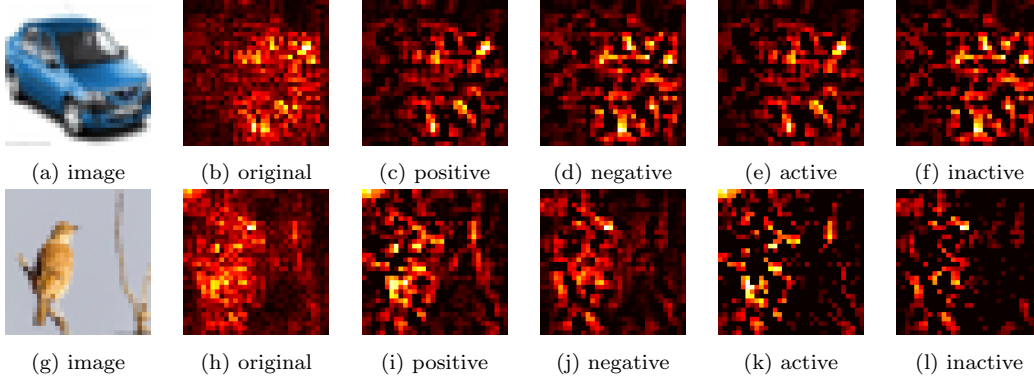(g) image   (h) original   (i) positive   (j) negative   (k) active   (l) inactive

Figure 1: Comparison Saliency Maps CNN Cifar10

4.1. Qualitative Evaluation

Following the same approach that can be found in the literature of Saliency Maps, first the techniques will be visually evaluated. In several papers, as in [5], [6], [10] or [7], the different techniques were compared based on a visual inspection.

One trend in the mentioned article is to just show some examples of the visualization to make the comparison. However, in none of them is explained how these examples are obtained. Therefore, it could be the case that the visualization is better for a certain class or example. To solve this problem in this paper a correct example for each class from the test is selected. These images are shown in Appendix Appendix A. Here, two examples can be observed in Figure 1 and Figure 2.

From the comparison it can be appreciated that the proposed techniques reduce the noise of the original Saliency Maps, producing sharper visualization. The shape of the objects that the neural network is classifying are is clearer and there is a higher level of detail.

(a) image   (b) original   (c) positive   (d) negative   (e) active   (f) inactive

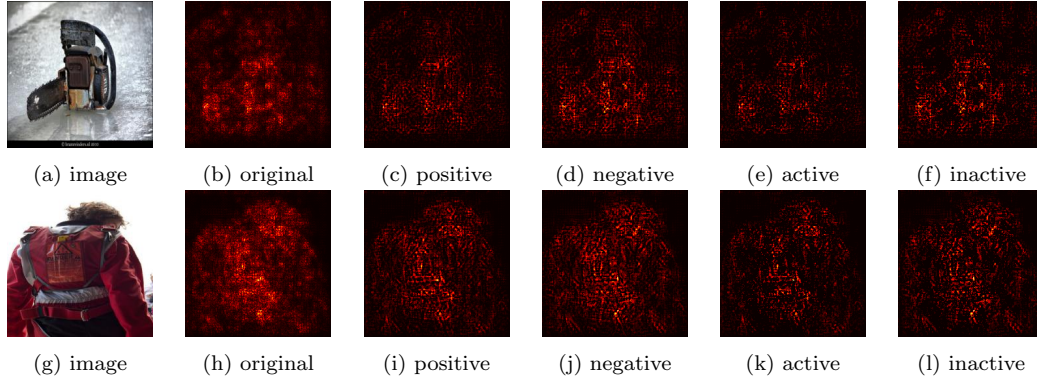(g) image   (h) original   (i) positive   (j) negative   (k) active   (l) inactive

Figure 2: Comparison Saliency Maps Resnet18 Imagenette
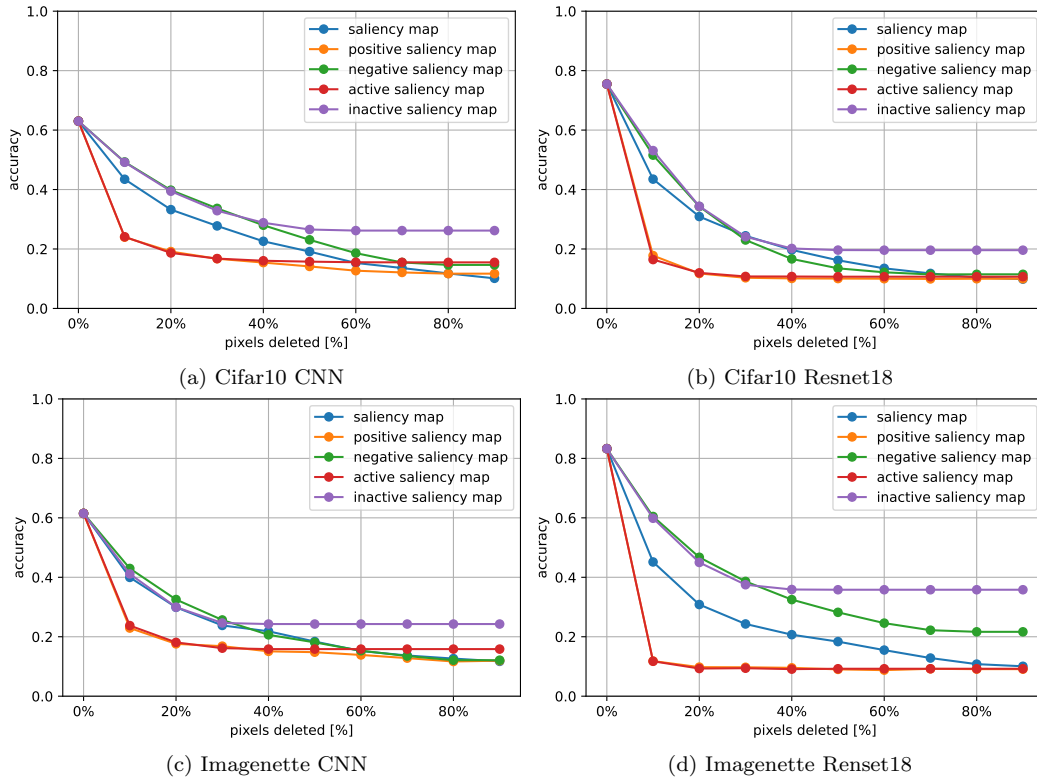

(a) Cifar10 CNN

(b) Cifar10 Resnet18

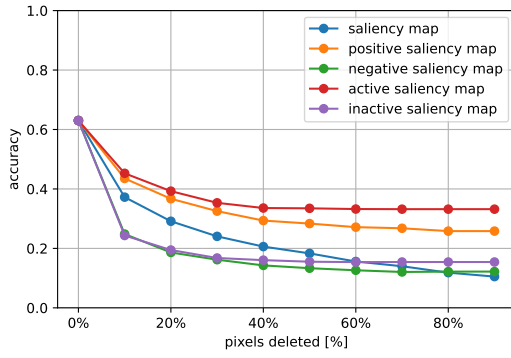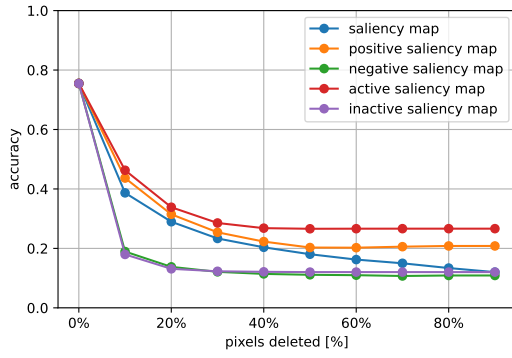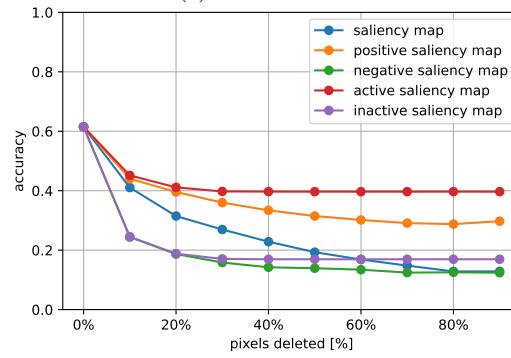(c) Imagenette CNN

(d) Imagenette Renset18
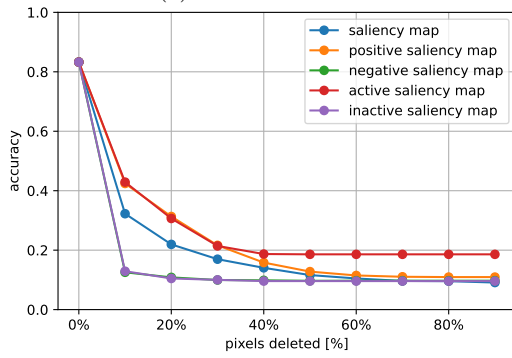
Figure 3: Black-Deletion Benchmark

(a) Cifar10 CNN

(b) Cifar10 Resnet18

(c) Imagenette CNN

(d) Imagenette Renset18

Figure 4: White-Deletion Benchmark

8

4.2. Quantitative Evaluation

Even though historically the evaluation of Saliency Maps has been quanitative, there have been efforts to try to formulate metrics to measure the effectiveness of these techniques, as in

References

[1] A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, in: Advances in Neural Information Processing Systems, Vol. 25, Curran Associates, Inc.
URL https://papers.nips.cc/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Al

[2] ImageNet.
URL https://www.image-net.org/

[3] G. Singh, Think positive: An interpretable neural network for image recognition 151 178–189. doi:10.1016/j.neunet.2022.03.034.
URL https://www.sciencedirect.com/science/article/pii/S0893608022001125

[4] Y. LeCun, Y. Bengio, T. B. Laboratories, Convolutional Networks for Images, Speech, and Time-Series 14.

[5] K. Simonyan, A. Vedaldi, A. Zisserman, Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. arXiv:1312.6034, doi:10.48550/arXiv.1312.6034.
URL http://arxiv.org/abs/1312.6034

[6] J. T. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller, Striving for Simplicity: The All Convolutional Net. arXiv:1412.6806, doi:10.48550/arXiv.1412.6806.
URL http://arxiv.org/abs/1412.6806

[7] M. Sundararajan, A. Taly, Q. Yan, Axiomatic Attribution for Deep Networks, in: Proceedings of the 34th International Conference on Machine Learning, PMLR, pp. 3319–3328.
URL https://proceedings.mlr.press/v70/sundararajan17a.html

[8] A. Shrikumar, P. Greenside, A. Shcherbina, A. Kundaje, Not Just a Black Box: Learning Important Features Through Propagating Activation Differences. arXiv:1605.01713, doi:10.48550/arXiv.1605.01713.
URL http://arxiv.org/abs/1605.01713

[9] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, B. Kim, Sanity Checks for Saliency Maps, in: Advances in Neural Information Processing Systems, Vol. 31, Curran Associates, Inc.
URL https://papers.nips.cc/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Al

[10] D. Smilkov, N. Thorat, B. Kim, F. Viégas, M. Wattenberg, SmoothGrad: Removing noise by adding noise 10.

[11] MNIST handwritten digit database, Yann LeCun, Corinna Cortes and Chris Burges.
URL http://yann.lecun.com/exdb/mnist/

[12] CIFAR-10 and CIFAR-100 datasets.
URL https://www.cs.toronto.edu/ kriz/cifar.html

[13] Imagenette.
URL https://github.com/fastai/imagenette

[14] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. doi:10.1109/CVPR.2016.90.

Appendix A. Signed Saliency Map Examples

hey this is gonna be awesome