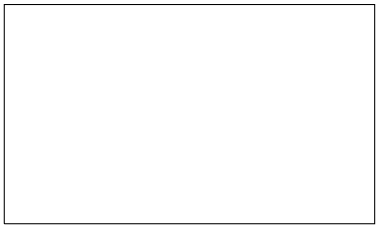
# Graphical Abstract

**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:

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

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

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, as far as we know, there is not any research about the meaning or impact of the sign in Saliency Maps. The only research that discuss briefly this aspect is [10]. In this paper it was explained that the raw value of gradients (without absolute value), it 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, that can be due to a high number of pixels with a value near zero in the Saliency Map: all the image would bright at a medium intensity. If that is the case, visualizing separately positive and negative gradients will produce clearer pictures. Because of this, in this paper two techniques are proposed:

- Positive Saliency Maps:

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

- Negative Saliency Maps:

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

## 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

## 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-A

[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-A

[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/`

## Appendix A. Sample Appendix Section

hey this is gonna be awesome