A matter of attitude: Focusing on positive and active gradients to boost 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: eXplainable Artificial Intelligence (XAI), Convolutional Neural Networks, Saliency maps, Visualization, Gradient signs

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 architectures.

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 explain these other models several techniques have been developed during the last years. One of the most used types are saliency maps.

Email addresses: `oscar.llorente.gonzalez@ericsson.com` (Oscar Llorente), `jaime.boal@iit.comillas.edu` (Jaime Boal), `eugenio.sanchez@iit.comillas.edu` (Eugenio F. Sánchez-Úbeda)

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:

$$\hat{c} = \underset{c \in C}{\operatorname{argmax}}\, S_c(\boldsymbol{I}), \tag{1}$$

being $\boldsymbol{I}$ the input image, C the set of all possible classes and S_c the classification score for a certain class c. Since usually an image is expressed in rgb format is important to note that $\boldsymbol{I} \in \mathbb{R}^{\text{channels}\,x\,\text{height}\,x\,\text{width}}$. Then, the original saliency map method is expressed as:

$$\boldsymbol{M}_{i\,j} = \max_{k} \left(\frac{\partial S_{\hat{c}}}{\partial \boldsymbol{I}_{k\,i\,j}} \right), \tag{2}$$

being $\boldsymbol{M}$ a 2D map, k an specific channel and, i and j, the height and width coordinates, respectively. 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. As it is shown in the previous equation, 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, 8]. By these, 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 using these techniques to improve the ones from the original saliency maps: more knowledge can be unlocked from the gradients of the neural network without further modifications. First, 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 a different meaning depending on whether 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 sets of pixels 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 darker, as for example an object in front of a black background. Based on this, two sets of pixels can be distinguished in the image:

- Pixels that by increasing their value, the classification score of the predicted class improves since they have a positive value in the saliency map (positive gradient).
- Pixels that by decreasing their value, the classification score of the predicted class 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 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:

$$\boldsymbol{M}_{ij} = \max_k \left(\mathrm{ReLU}\left(\frac{\partial S_{\hat{c}}}{\partial \boldsymbol{I}_{kij}} \right) \right), \tag{3}$$

- Negative saliency maps:

$$\boldsymbol{M}_{ij} = \max_k \left(\mathrm{ReLU}\left(-\frac{\partial S_{\hat{c}}}{\partial \boldsymbol{I}_{kij}} \right) \right), \tag{4}$$

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, from the derivative respect to other class score, with a higher value than the one respect to the predicted class and the current techniques cannot show that to the user. If it is the case, and the value of this pixel is increased, the classification score for the other class will improve more than the one from the original class, making the prediction of the original class worse, instead of better. 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 that a gradient is higher or lower than another means.

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 (positive and negative saliency maps):

- Pixels that by increasing their value the classification score for the predicted class will improve more that the score from any other class. Since now other classes are taken into account, only the pixels that have a higher gradient with respect to the predicted class than with respect to any other class, will improve the classification if their value is increased.
- Analogously, the pixels that by decreasing their value classification score for the predicted class will improve more that the score from any other class, are only the ones that have a lower gradient with respect to the predicted class than with respect to 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 score from the original prediction):

$$\boldsymbol{M}_{ij} = \max_k \begin{cases} \frac{\partial S_{\hat{c}}}{\partial \boldsymbol{I}_{kij}}, & \frac{\partial S_{\hat{c}}}{\partial \boldsymbol{I}_{kij}} = \mathrm{argmax}_{c \in C} \frac{\partial S_c}{\partial \boldsymbol{I}_{kij}} \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

- Inactive saliency map (decreasing them improves the classification score from the original prediction):

$$\boldsymbol{M}_{ij} = \max_k \begin{cases} \frac{\partial S_{\hat{c}}}{\partial \boldsymbol{I}_{kij}}, & \frac{\partial S_{\hat{c}}}{\partial \boldsymbol{I}_{kij}} = \mathrm{argmin}_{c \in C} \frac{\partial S_c}{\partial \boldsymbol{I}_{kij}} \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

Note that, both type of methods, positive and negative saliency maps, and active and inactive saliency maps, pursue the same purpose. Positive and active try to find the pixels that by being brighter, increase the confidence of the original classification. Analogously, negative and inactive try to find the pixels that by being darker, also increase the confidence of the original classification.

Positive and negative take into account the sign, which is a significant improvement from the original saliency maps. Active and inactive can be seen as a step further, taking into account also the other classes. Even so, positive and negative can also provide information about which pixels increase the score of the predicted class, even though another score is increasing more. This information can also help the user understand how the class scores of the model works.

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, CIFAR-10 [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, for CIFAR-10 two model will be used. The first is a CNN model based in a max-pooling, several layers for CNNs, an average-pooling and a final linear layer. The second one is the classical ResNet-18 [14]. Both model will be trained from scratch. On the other hand, for Imagenette, in addittion to these two models, a pre-trained version of the ResNet-18 and of the ConvNeXt [15] will be used. 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:

Table 1: Test set accuracy.

	CIFAR-10	Imagenette
Basic CNN	0.6307	0.6205
ResNet-18	0.7569	0.8357
ResNet-18 pre-trained	-	0.9733
ConvNext pre-trained	-	0.9932

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.

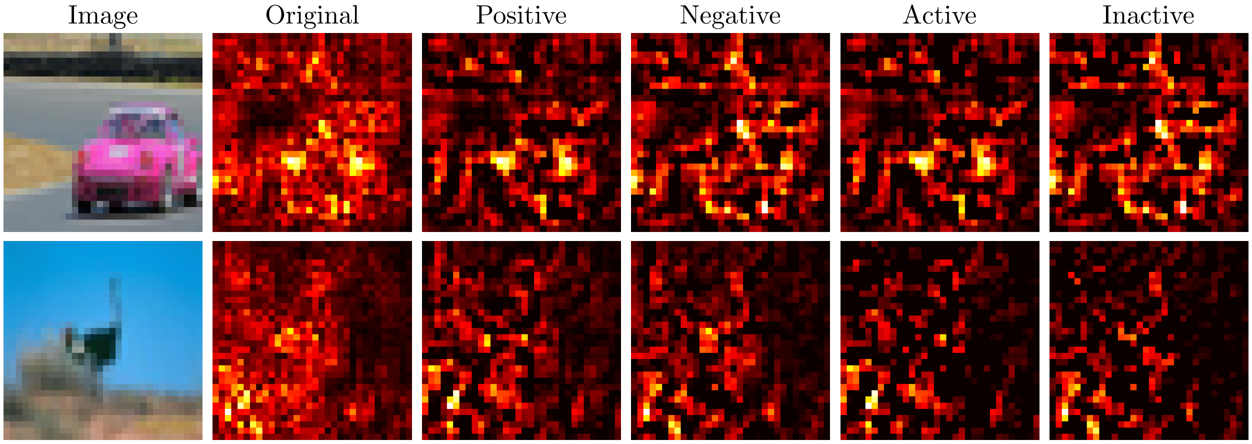


Figure 1: Comparison saliency maps CNN CIFAR-10.

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 randomly selected. These images are shown in 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

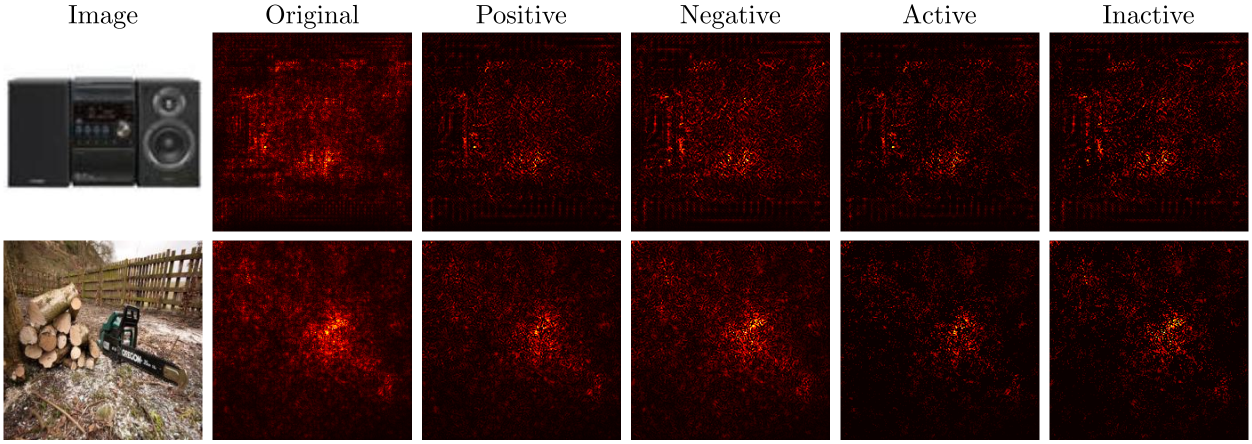


Figure 2: Comparison saliency maps ResNet-18 Imagenette.

that the neural network is classifying are is clearer and there is a higher level of detail. However, and even though this evaluation is the common approach in the literature, it does not prove that sharper or less noisy visualizations are better. It could be the case that a noisier visualization is more faithful to what the neural network has learned. Because of this, a quantitative evaluation has been done.

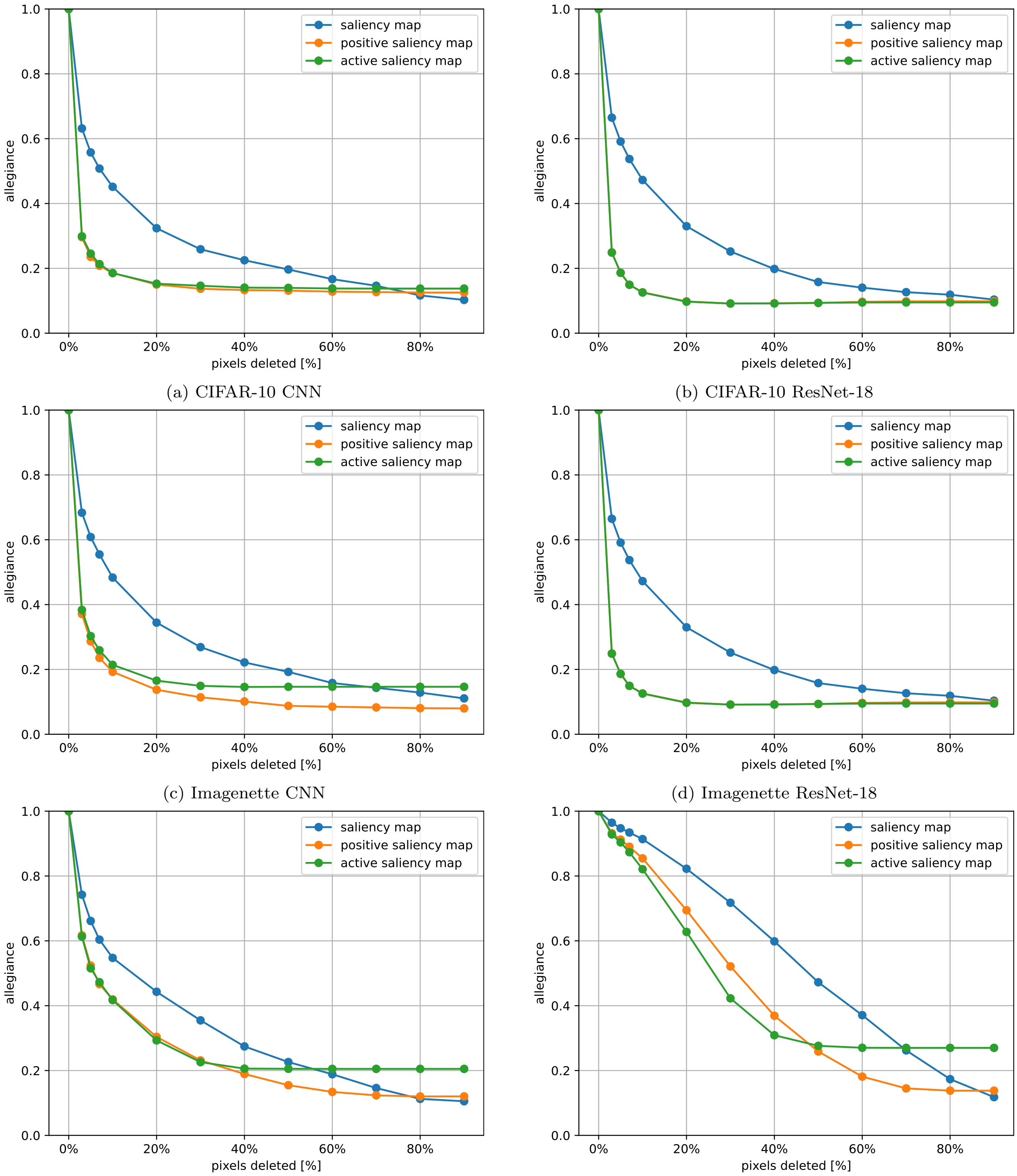
4.2. Quantitative evaluation

Even though historically the evaluation of saliency maps has been qualitative, there have been efforts to try to formulate metrics to measure the effectiveness of these techniques, as in [16], [17] or [18]. While in [18] it is studied a desirable property of explainability techniques, in [16] and [17] the objective is to formulate a metric that could indicate which technique is better.

Specifically, [16] proposed a metric called deletion that deletes pixels in descending order of importance (according to the explainability technique) and recomputes the probability of the correct output for each fraction of deleted pixels. For the mentioned deletion researchers fixed the values of these pixels to a constant, as black or gray, or to random noise.

In [17] it was proposed a variation of the deletion metric. In this paper it was explained that it was necessary to retrain the model for each portion of deleted pixels to maintain the same distribution in the training set and the test set, where the explainability technique was being measured. However, in the retraining, the model can learn new other things, so the metric is not measuring how well the explainability technique represents what the neural network has learned because the information used by the model to classify can be different now.

Going back to the deletion metric, its problem is that the value of a pixel cannot be deleted, because for a neural network any value can have some meaning, including the black, the gray or noise. Then, the value introduced can affect different pixels in different ways, introducing unknown biases. However, the separation in two different sets of pixels, the ones that should be brighter (positive and active saliency maps) to improve the classification score of the original predictions and the ones that should be darker (negative and inactive saliency maps), opens the door to implement the deletion technique knowing what the introduced



(a) CIFAR-10 CNN

(b) CIFAR-10 ResNet-18

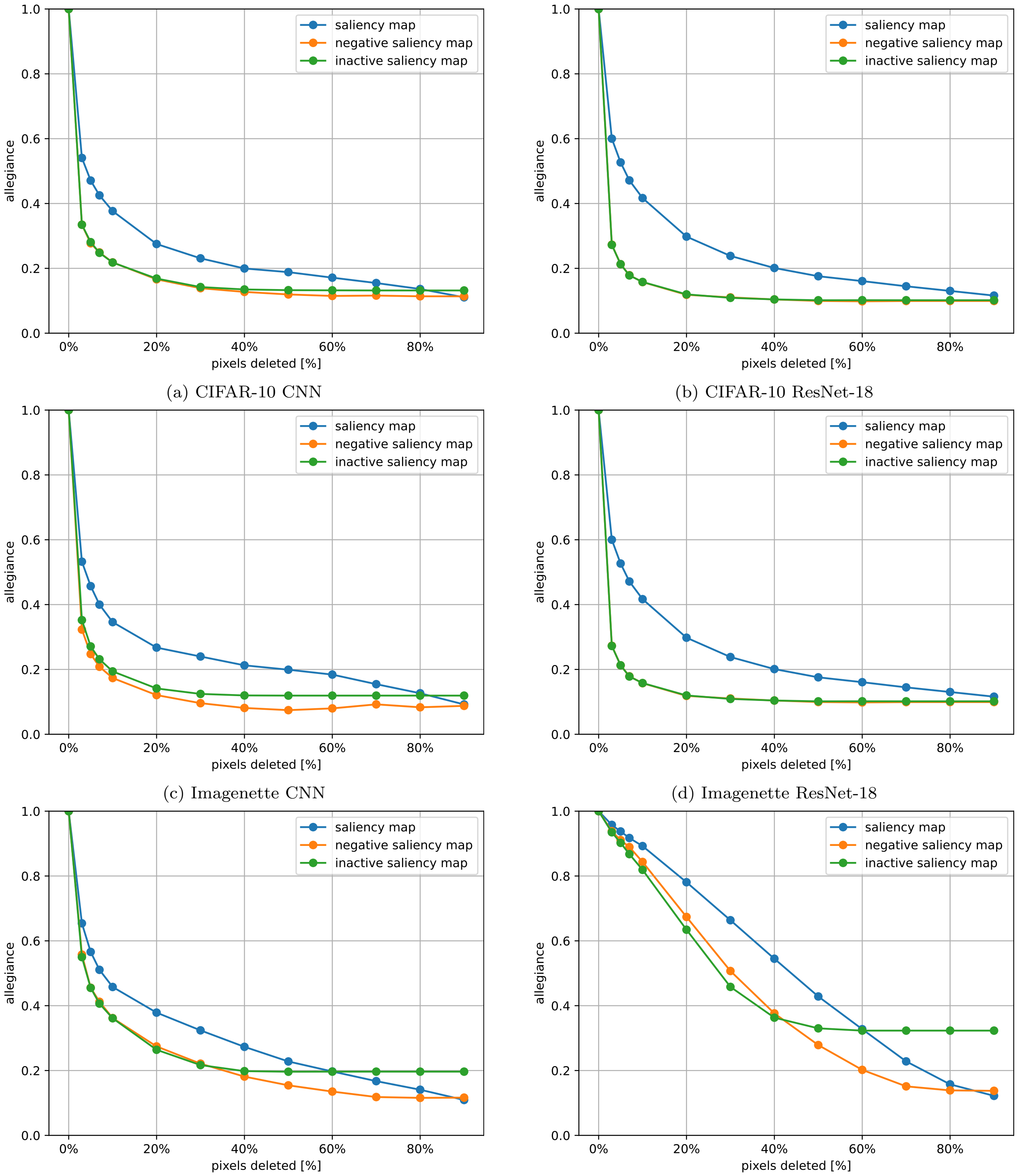
(c) Imagenette CNN

(d) Imagenette ResNet-18

(e) Imagenette ResNet-18 pre-trained

(f) Imagenette ConvNeXt pre-trained

Figure 3: Black-Deletion Benchmark.



(a) CIFAR-10 CNN

(b) CIFAR-10 ResNet-18

(c) Imagenette CNN

(d) Imagenette ResNet-18

(e) Imagenette ResNet-18 pre-trained

(f) Imagenette ConvNeXt pre-trained

Figure 4: White-Deletion Benchmark.

Table 2: AUC for black deletions in saliency maps.

	CIFAR-10		Imagenette			
	CNN	ResNet-18	CNN	ResNet-18	ResNet-18 pre-trained	ConvNeXt
Original	0.23	0.23	0.24	0.26	0.28	0.49
Positive	0.14	0.11	0.12	0.14	0.21	0.37
Active	0.15	0.11	0.16	0.14	0.24	0.39

Table 3: AUC for white deletions in saliency maps.

	CIFAR-10		Imagenette			
	CNN	ResNet-18	CNN	ResNet-18	ResNet-18 pre-trained	ConvNeXt
Original	0.21	0.22	0.21	0.23	0.26	0.46
Negative	0.14	0.12	0.11	0.18	0.20	0.37
Inactive	0.15	0.12	0.14	0.16	0.23	0.42

value means. In this paper the two proposed metrics are called black-deletion and white-deletion.

First, the new techniques give the user the two different sets of pixels (should be brighter and should be darker). Since increasing the value (making them brighter) of some of them improves the classification score of the predicted classes, putting them to zero (black-deletion) should harm severely the original classification. Analogously, putting the pixels that should be darker to improve the original classification to one (white deletion) should show also harm severely the original classification. Therefore, black-deletion will help to measure the effect of positive and active salience maps, while white deletion will be used to measure negative and inactive salience maps. Both metrics will measure the change in the predicted classes with respect to the original classification (something that here has been called allegiance) for the following fractions of deleted pixels: [0.03, 0.05, 0.07, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]. The interval to measure the deletion is 10%, since based on the literature [17] it a common interval, with a higher level of detail at the beginning to show the sharp decrease in the first deletions. These can be appreciated in Figure 3 and Figure 4.

In the graphs it can be seen that the behavior is the expected, meaning that the selection of positive, negative, active and inactive gradients worked. As expected, the decrease in allegiance for black-deletion is greater for active and positive than for the original saliency map. Analogously, the decrease in allegiance for white-deletion is greater for inactive and negative saliency maps than for the original ones. This proves that the pixels identified.

An important thing to note is that in the graphs it can be seen that for active and inactive saliency maps there is almost no change after 50% of the pixels have been deleted. The reason for this is that in the active and inactive there are a lot of pixels that are black (because its derivative with respect to the original predicted class is not greater (active) or lower (inactive) than for any other class). Therefore, after all the pixels from the active and inactive saliency maps are deleted there is no change since the pixels that have zero value in the saliency maps were ignored in the deletion (deleting a pixel that does not belong to the saliency map variant does not give us any information). The same happens for the positive and negative later, around 80%, showing that there are more pixels with positive and negative derivatives, than pixels with a derivative higher or lower with respect to the original class than for any other class.

To give concrete numbers, the area under the curve was also calculated and showed in Table 2 for black-deletion and in Table 3 for white-deletion. The intuition of what would happen in the experiments was correct and it has been shown that the new proposed method help the user in identifying pixels that, by being brighter or darker, help the confidence of original predicted classes. However, although the improvement over the original saliency map is clear, it is surprising how positive and negative sometimes work better than active and inactive. The main reason for this could be that these metrics are still not perfect, since the value used for the deletion are the extreme values, and pixels that may improve the classification by being brighter or darker, but that it is not the same that putting their value to white or black. However, from all the possible values to choose, we find most reasonable use the darkest and the brightest value since any pixel can be darker and brighter than them, respectively. Even so, it is still better than the other metrics use in the literature since it introduces a know alteration in the image instead of an unknown one.

Note that every computation shown in this section is based on the test dataset. To see the ones from the training, that have similar results, go to Appendix Appendix B.

5. Conclusion and future work

To sum up, in this paper it has been shown that there is more information hidden in the gradients, both in the sign of the gradients and in the gradients of other classes. This paper is the first study that takes into account the sign and the other classes in the explainability, which seem to be crucial for extracting all the information contained in saliency maps. Therefore, it is not necessary to create complicated techniques to create a sharper visualization that shows what the model is looking when it is classifying. Moreover, this can be the starting point of a reformulation of all saliency variants, since it can produce improvements in all the techniques that are based on the gradients. However, this is out of the scope of this article.

Another contribution worth of mentioning is that taking into account the sign or the other classes allows comparing different techniques with the proposed metric, black-deletion and white-deletion. These metrics are the first benchmark that introduces an alteration knowing its meaning, instead of introducing an unknown bias. Therefore, they can also be used to compare more faithfully which are the best saliency techniques. We leave these for future research.

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., 2012.

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

[3] G. Singh, Think positive: An interpretable neural network for image recognition, Neural Networks 151 (2022) 178–189. doi:10.1016/j.neunet.2022.03.034.

[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 (Apr. 2014). arXiv:1312.6034, doi:10.48550/arXiv.1312.6034.

[6] J. T. Springenberg, A. Dosovitskiy, T. Brox, M. Riedmiller, Striving for Simplicity: The All Convolutional Net (Apr. 2015). arXiv:1412.6806, doi:10.48550/arXiv.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, 2017, pp. 3319–3328.

[8] A. Shrikumar, P. Greenside, A. Shcherbina, A. Kundaje, Not Just a Black Box: Learning Important Features Through Propagating Activation Differences (Apr. 2017). arXiv:1605.01713, doi:10.48550/arXiv.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., 2018.

[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, http://yann.lecun.com/exdb/mnist/.

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

[13] Imagenette, fast.ai (Sep. 2022).

[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), 2016, pp. 770–778. doi:10.1109/CVPR.2016.90.

[15] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A ConvNet for the 2020s (Mar. 2022). arXiv:2201.03545, doi:10.48550/arXiv.2201.03545.

[16] V. Petsiuk, RISE: Randomized Input Sampling for Explanation of Black-box Models 13.

[17] S. Hooker, D. Erhan, P.-J. Kindermans, B. Kim, A Benchmark for Interpretability Methods in Deep Neural Networks, in: Advances in Neural Information Processing Systems, Vol. 32, Curran Associates, Inc., 2019.

[18] M. Ancona, E. Ceolini, C. Öztireli, M. Gross, Towards better understanding of gradient-based attribution methods for Deep Neural Networks, in: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings, OpenReview.net, 2018.

Appendix A. Signed saliency map examples

In this appendix the examples for each technique will be showed for all the combinations:

Appendix B. Black-Deletion and White-Deletion for Trainning set