Iteratively unveiling new regions of interest in Deep Learning models

Florian Bordes[1] **Tess Berthier**[2] **Lisa Di Jorio**[2] **Pascal Vincent**[1,3] **Yoshua Bengio**[1,3]
[1]Montréal Institute for Learning Algorithms (MILA), Université de Montréal
[2]Imagia Cybernetics [3]CIFAR

Abstract

Recent advance of deep learning has been transforming the landscape in many domains. However, understanding the predictions of a deep network remains a challenge, which is especially sensitive in health care domains as interpretability is key. Techniques that rely on saliency maps *-highlighting the region of an image that influence the classifier's decision the most-* are often used for that purpose. However, gradients fluctuation make saliency maps noisy and thus difficult to interpret at a human level. Moreover, models tend to focus on one particular influential region of interest (ROI) in the image, even though other regions might be relevant for the decision.
We propose a new framework that refines those saliency maps to generate segmentation masks over the ROI on the initial image. In a second contribution, we propose to apply those masks over the original inputs, then evaluate our classifier on the masked inputs to identify previously overlooked ROI. This iterative procedure allows us to emphasize new region of interests by extracting meaningful information from the saliency maps.

1 Introduction and motivation

In health care domains, medical experts need to understand and validate machine learning prediction. Due to their lack of interpretability, deep neural networks are often qualified as black box and are not easily trusted by clinicians. Moreover, as they are used for segmentation or localization, these models often need auxiliary mechanisms that make them more expensive to train.

More computationally efficient approaches emphasizing regions of interest (ROI) for deep net classifiers have been investigated. A first one called activation maximization [1] perform gradient ascent in the input space in order to maximize a unit activation. Simonyan et al. show that DeconvNet [10] is equivalent to a gradient back-propagation through a ConvNet and present saliency maps which highlight ROI. Another interesting approach is SmoothGrad [7], which averages the vanilla saliency maps of n noisy images resulting in a more visually coherent map. In recent works, Kindermans et al. question the use of the gradient's direction to estimate signal in the data and propose new methods to take the data distribution into account.

However those algorithms show that classifiers tend to focus on the most influential region of interest (ROI), sometimes overlooking other ROIs that might have been relevant for the decision. Moreover, improving the visual coherence of these saliency maps is still key to more interpretable models. This calls for new frameworks that highlight, not only one, but all the regions in an image contributing to any classifier's decision while keeping coherency.

In this paper we use guided backpropagation and smoothgrad algorithms to obtain our saliency maps. For the purpose of improving the visual coherence of the maps, we suggest smoothing them by iterating a combination of convolutions and thresholds. This results in smoothed saliency maps that

1st Conference on Medical Imaging with Deep Learning (MIDL 2018), Amsterdam, The Netherlands.

coherently replicate the ROI shape. Those new maps will serve as masks and superimpose their corresponding ROI, thus localizing them in the input image.

We propose an iterative procedure where predictions are run at each step with images that are more and more masked. Feeding our newly masked input to the classifier will force it to focus on a ROI previously overlooked.

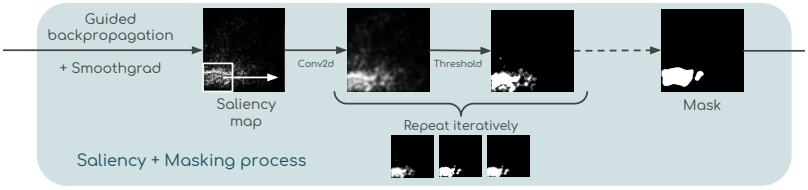

Figure 1: Segmentation mask process on a saliency map from ChestX-ray8.

2 Proposed approach

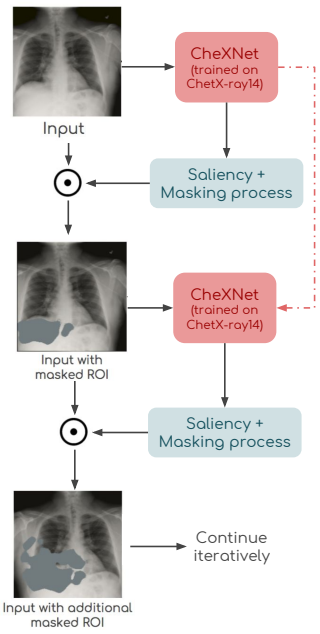

Figure 2: Architecture of the iterative process, unveiling new ROI.

Saliency maps, also referred as activation maps, tend to reflect the influence of an individual pixel in the final classification. Saliency map can be simply generated by computing the gradient of the maximally responding output unit with respect to the input [1] [6]. As the saliency maps based on this raw gradients are visually noisy, we preferred to implement a guided backpropagation algorithm [8] with SmoothGrad [7] resulting in more visually coherent maps.

Several techniques could be used to smooth saliency maps such as algorithms based on random markov field. However it would require to sample many times for each pixel which is computationally expensive. Instead, we designed a convolution based algorithm that computes iteratively a score (mean over a neighborhood of pixels) and assigns a threshold for those values. If a pixel has the same value than its neighbors it will remain the same, and it will take the value of the other neighbors otherwise. In order to smooth all the points, we need to repeat this procedure several times depending on the filter and image size as shown in Figure 1.

Once the masks are obtained, we apply them over the inputs, consequently hiding the ROI previously found by the classifier. This new image is then fed to our neural network which will evaluate it to produce a new mask. We can repeat this procedure until the class predicted probability is below a given threshold. (Figure 2).

Most of the parameters used in this method depend on the dataset and the degree of precision wanted in the segmentation. For example, a bigger kernel and number of iterations for the convolution result in a bigger mask that can help to localize a ROI in the image. In opposite, a small number of iterations and kernel size result in a more precise mask that could locate small lesions or tumors.

3 Experiments

We started experimenting with a pretrained resnet model [2] on the Standford dog dataset [3], which is often used for fine-grained classification. We then used CheXNet [5], a deep model trained on the public dataset ChestX-ray14 [9]. For each experiment we obtained our saliency maps using the smooth Grad algorithm. We convolved three times the saliency maps with filter of ones with size eleven by eleven in order to get the segmentation mask. After applying the segmentation mask over the inputs, we repeated the previous procedure 5 times. Results are presented in Figure 3 where

we can observe the segmentation masks being constructed iteratively for both examples. In those experiments, we did not train nor fine-tune any model in between each new masking process.

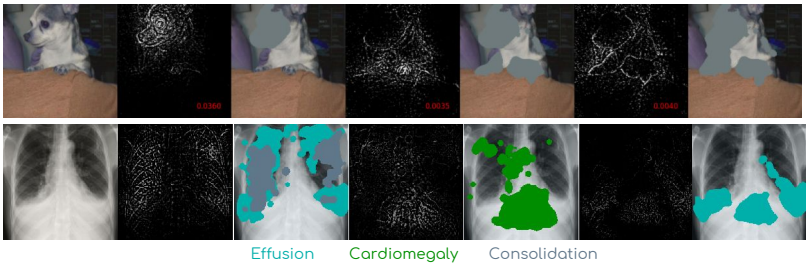

Figure 3: Examples of our procedure on ImageNet and ChestXRay

As we can see in the first example of figure 3, the neural network will first look for the dog's face in order to classify. However after masking the face, the model is starting to use the paws as the next ROI even if they were not really visible in the previous saliency map. In the second example (ChestXray), at first the CNN classifies the effusion and consolidation only, moreover the effusion covers the whole lung. After one iteration the cardiomegaly (enlarged heart) is detectable by the classifier. The second iteration gives a better localization of the effusion (bottom of the lungs).

4 Future Work

In order to make our iterative procedure more effective, we are planning to train models on the masked inputs between each iteration. However this method could raise new questions, such as the classifier learning how to classify masks and interpreting them. Further improvements in order to produce better saliency maps could be made, such as implementing methods like PatterNet [4].

References

[1] Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. Technical Report 1341, University of Montreal, June 2009.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[3] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[4] Pieter-Jan Kindermans et al. Learning how to explain neural networks: Patternnet and patternattribution. arxiv preprint. *arXiv preprint arXiv:1705.05598*, 2017.

[5] Pranav Rajpurkar et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[6] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[7] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[8] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for simplicity: The all convolutional net. *CoRR*, abs/1412.6806, 2014.

[9] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. IEEE, 2017.

[10] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.