Opti-CAM: Optimizing saliency maps for interpretability

Hanwei Zhang, Felipe Torres, Ronan Sicre, Stephane Ayache and Yannis Avrithis

Centrale Marseille, Aix Marseille Univ, CNRS, LIS, Marseille, France.

https://arxiv.org/abs/2301.07002

LABORATOIRE D'INFORMATIQUE & SYSTÈMES

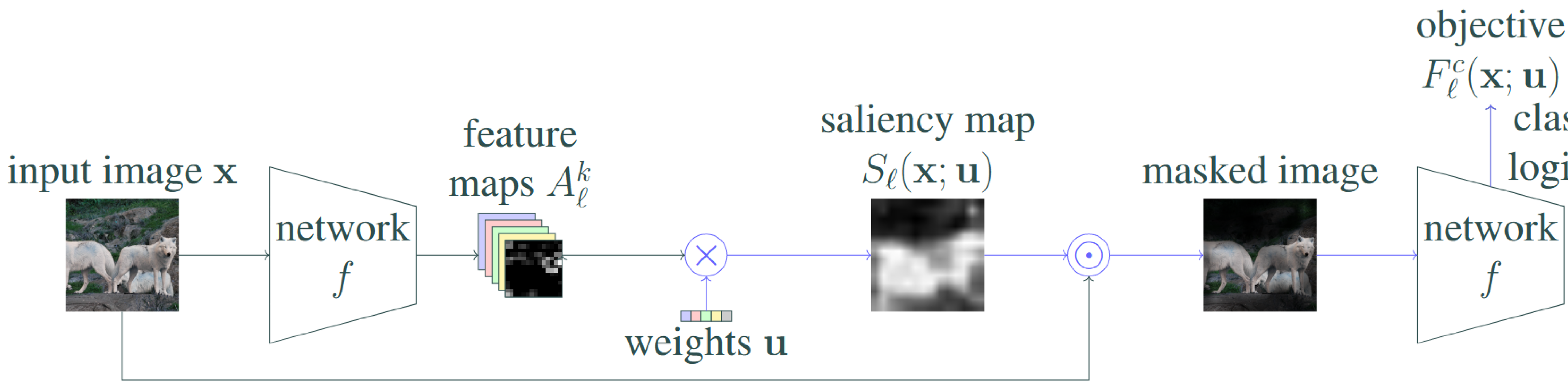

Figure 1: Overview of Opti-CAM. Given an input image $\mathbf{x}$, a fixed network f, a target layer ℓ and a class of interest c, we extract the feature maps from layer ℓ and obtain a saliency map $S_\ell(\mathbf{x};\mathbf{u})$ by combining the feature maps ($\times$) with weights from variable $\mathbf{u}$ (5). After upsampling and normalizing, the saliency map is element-wise multiplied ($\odot$) with the input image and fed to f. We find $\mathbf{u}^*$ maximizing $F_\ell^c(\mathbf{x};\mathbf{u})$ along the path highlighted in blue.

Abstract

Methods based on *class activation maps* (CAM) interpret predictions of Deep neural networks (DNN) by using a linear combinations of feature maps as saliency maps. By contrast, masking-based methods optimize a saliency map directly in the image space or train another network on additional data to build it.

We introduce Opti-CAM, combining ideas from CAM-based and masking-based approaches. Our saliency map is a linear combination of feature maps, where weights are optimized per image such that the logit of the masked image for a given class is maximized. We also study evaluation metrics and propose the Average Gain.Opti-CAM largely outperforms other CAM-based approaches. We also show that localization and classifier interpretability are not necessarily aligned.

Background

CAM-based saliency maps are built as a linear combination of feature maps $A_\ell^k = f_\ell^k(\mathbf{x})$. For layer ℓ and class c, the saliency is

$$S_\ell^c(\mathbf{x}) := h\left(\sum_k w_k^c A_\ell^k\right), \tag{1}$$

where w_k^c are the weights of each channel and h an activation function.

Grad-CAM is defined with $h = \mathrm{relu}$ and weights

$$w_k^c := \mathrm{GAP}\left(\frac{\partial y_c}{\partial A_\ell^k}\right), \tag{2}$$

where GAP is global average pooling.

Score-CAM [2] is defined with $h = \mathrm{relu}$ and weights $w_k^c := \mathrm{softmax}(\mathbf{u}^c)_k$, where $\mathbf{u}^c$ is the increase in confidence for class c of the input image $\mathbf{x}$ masked by the saliency map:

$$u_k^c := f(\mathbf{x} \odot n(\mathrm{up}(A_\ell^k)))_c - f(\mathbf{x}_b)_c, \tag{3}$$

where $\odot$ is the Hadamard product, up is upsampling and n the saliency map normalization.

Masking-based methods rely on optimization in the input space, like *extremal perturbations* [1]. Optimization often takes the form

$$S^c(\mathbf{x}) := \arg\max_{\mathbf{m}\in\mathcal{M}} f(\mathbf{x} \odot n(\mathrm{up}(\mathbf{m})))_c + \lambda R(\mathbf{m}). \tag{4}$$

Here, a mask $\mathbf{m}$ is directly optimized and does not rely on feature maps of any layer. However, the optimization is complex and requires regularization.

Opti-CAM

As CAM methods, our saliency map is a combination of feature maps, but we optimize the weights given an objective function. We use channel weights $w_k := \mathrm{softmax}(\mathbf{u})_k$, where $\mathbf{u}$ is the variable. Our saliency map S_ℓ is a function of input $\mathbf{x}$ and variable $\mathbf{u}$:

$$S_\ell(\mathbf{x};\mathbf{u}) := \sum_k \mathrm{softmax}(\mathbf{u})_k A_\ell^k. \tag{5}$$

Given a layer ℓ, we find the vector $\mathbf{u}^*$ that maximizes the classifier confidence for class c, when the input image $\mathbf{x}$ is masked according to saliency map $S_\ell(\mathbf{x};\mathbf{u}^*)$:

$$\mathbf{u}^* := \arg\max_{\mathbf{u}} F_\ell^c(\mathbf{x};\mathbf{u}), \quad \text{where} \quad F_\ell^c(\mathbf{x};\mathbf{u}) := g_c(f(\mathbf{x} \odot n(\mathrm{up}(S_\ell(\mathbf{x};\mathbf{u}))))). \tag{6}$$

The saliency map $S_\ell(\mathbf{x};\mathbf{u})$ is adapted to $\mathbf{x}$ by upscaling and normalizing. Finally we have

$$S_\ell^c(\mathbf{x}) := S_\ell(\mathbf{x};\mathbf{u}^*) = S_\ell(\mathbf{x};\arg\max_{\mathbf{u}} F_\ell^c(\mathbf{x};\mathbf{u})), \tag{7}$$

Figure 1 shows Opti-CAM, without details like upsampling and normalization. Optimization takes place along the highlighted path from variable $\mathbf{u}$ to objective function F_ℓ^c.

Results

Visualization of saliency maps on ImageNet and medical data are given in Figure 5.

Classification metrics: average drop/increase (AD, AI) measure the increase/drop of prediction when masking the input image with the saliency map. Since a trivial solution Fake-CAM exist, we propose to complete them with average gain (AG), see Figure 2.

Insertion (I) and deletion (D) iteratively insert/delete pixels from the input image and measure its impact on prediction, but these metrics favour small, compact saliency maps.

Localization metrics are often used to evaluate saliency maps, see Figure 3, but a network decision does not only take the object into account but the context as well. We show how bounding box, and background perform, when used as saliency map, see Figure 4.

METHOD	RESNET50			VGG16			VIT-B			DEIT-B			RESNET50		VGG16	
	AD↓	AG↑	AI↑	AD↓	AG↑	AI↑	AD↓	AG↑	AI↑	AD↓	AG↑	AI↑	I↑	D↓	I↑	D↓
Fake-CAM	0.8	1.6	46.0	0.5	0.6	42.6	0.3	0.4	48.3	0.6	0.3	44.6	50.7	28.1	46.1	26.9
Grad-CAM	12.2	17.6	44.4	14.2	14.7	40.6	69.4	2.5	12.4	33.5	1.7	12.5	66.3	14.7	**64.1**	11.6
Grad-CAM++	12.9	16.0	42.1	17.1	10.2	33.4	86.3	1.5	1.0	50.7	0.9	7.2	66.0	14.7	62.9	12.2
Score-CAM [2]	8.6	26.6	56.7	13.5	15.6	41.7	32.0	6.2	33.0	53.6	2.2	12.2	65.7	16.3	62.5	12.1
XGrad-CAM	12.2	17.6	44.4	13.8	14.8	41.2	88.1	0.4	4.3	80.5	0.3	4.1	66.3	14.7	**64.1**	11.7
Layer-CAM	15.6	15.0	38.8	48.9	3.1	13.5	82.0	0.2	2.9	88.9	0.4	2.6	67.0	**14.2**	58.3	**6.4**
ExPerturbation [1]	38.1	9.5	22.5	43.0	7.1	20.5	28.8	6.2	24.4	60.9	2.0	8.5	**70.7**	15.0	61.1	15.0
Opti-CAM (ours)	**1.5**	**68.8**	**92.8**	**1.3**	**71.2**	**92.7**	**0.6**	**18.0**	**90.1**	**0.9**	**26.0**	**83.5**	62.0	19.7	59.2	11.0

Figure 2: *Classification metrics* on ImageNet validation set, using CNNs and Transformers. AD/AI/AG: average drop/increase/gain; I/D: insertion/deletion; bold: best, excluding Fake-CAM.

METHOD	RESNET50							VGG16						
	OM↓	LE↓	F1↑	BA↑	SP↑	EP↑	SM↓	OM↓	LE↓	F1↑	BA↑	SP↑	EP↑	SM↓
Fake-CAM	63.6	54.0	57.7	47.9	99.8	28.5	0.98	64.7	54.0	57.7	47.9	99.8	28.5	1.07
Grad-CAM	72.9	65.8	49.8	**56.2**	69.8	33.3	1.30	71.1	62.3	42.0	54.2	64.8	32.0	1.39
Grad-CAM++	73.1	66.1	**50.4**	**56.2**	69.9	33.1	1.29	70.8	61.9	44.3	55.2	66.2	32.3	1.38
Score-CAM [2]	**72.2**	64.9	49.6	54.5	68.7	32.4	**1.25**	71.2	62.5	**45.3**	**58.5**	**68.2**	33.4	1.40
Ablation-CAM	72.8	65.7	50.2	56.1	69.9	33.1	1.26	71.3	62.6	43.2	56.2	65.7	32.7	1.39
XGrad-CAM	72.9	65.8	49.8	**56.2**	69.8	33.3	1.30	70.8	62.0	41.9	53.5	64.4	31.6	1.41
Layer-CAM	73.1	66.0	50.1	55.5	**70.0**	33.0	1.29	70.5	61.5	28.0	54.7	65.0	32.4	1.45
ExPerturbation [1]	73.6	66.6	37.5	44.2	64.8	**38.2**	1.59	74.1	66.4	37.8	43.3	62.7	**36.1**	1.74
Opti-CAM (ours)	**72.2**	**64.8**	47.3	49.2	59.4	30.5	1.34	**69.1**	**59.9**	44.1	51.2	61.4	30.7	**1.34**

Figure 3: *Localization metrics* on ImageNet. OM: *official metric*; LE: *localization error*; F1: *pixel-wise F_1 score*; BA: box accuracy; SP: standard pointing game; EP: energy pointing game; SM: *saliency metric*.

METHOD	AD↓			AG↑			AI↑		
	S	$B\cap S$	$S\setminus B$	S	$B\cap S$	$S\setminus B$	S	$B\cap S$	$S\setminus B$
$S := B$	67.2	–	–	2.3	–	–	9.2	–	–
$S := I \setminus B$	44.0	–	–	2.8	–	–	16.3	–	–
Fake-CAM	0.5	67.2	44.1	0.7	2.3	2.8	42.0	9.2	18.9
Grad-CAM	15.0	72.6	52.1	15.3	1.8	6.0	40.4	8.4	19.4
Grad-CAM++	16.5	72.9	53.1	10.6	1.6	4.1	35.2	7.3	17.1
Score-CAM [2]	12.5	71.5	50.5	16.1	2.2	6.3	42.5	8.6	20.8
Ablation-CAM	15.1	72.8	52.1	13.5	1.7	5.6	39.9	7.8	19.0
XGrad-CAM	14.3	72.6	51.4	15.1	1.8	6.0	42.1	8.0	20.1
Layer-CAM	49.2	84.2	74.4	2.7	0.4	1.2	12.7	4.4	7.3
ExPerturbation [1]	43.8	81.6	71.0	7.1	1.4	3.2	18.9	5.6	11.1
Opti-CAM (ours)	**1.4**	**62.5**	**34.8**	**66.3**	**8.7**	**25.8**	**92.5**	**18.6**	**47.1**

Figure 4: *Bounding box* study. Classification metrics on ImageNet using VGG16. B: ground-truth box used by localization metrics; I: entire image; S: saliency map. Bold: best, excluding Fake-CAM.

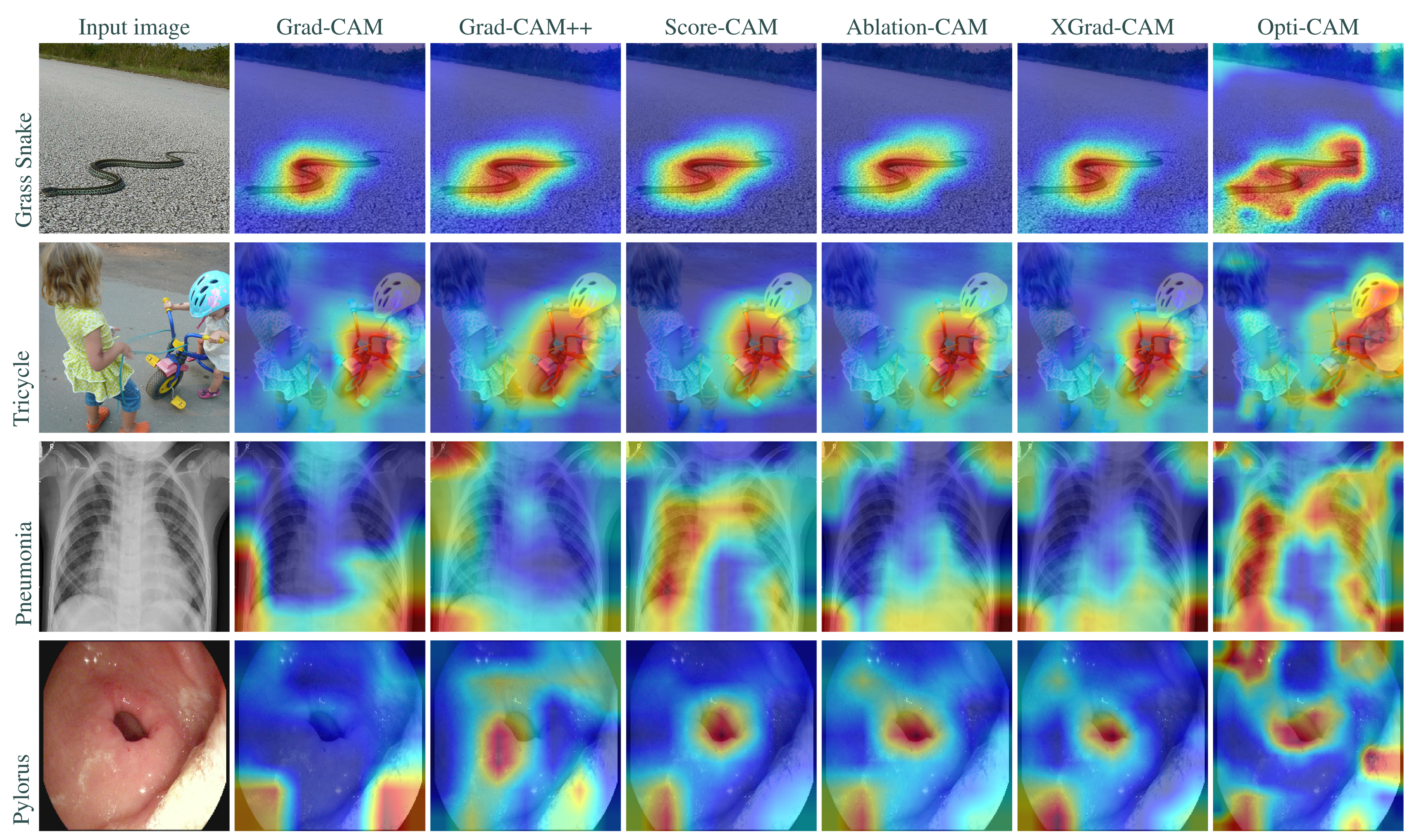

Figure 5: Saliency maps obtained on ImageNet (top two rows), Chest X-ray and Kvasir with VGG16.

References

[1] R. Fong, M. Patrick, and A. Vedaldi. Understanding deep networks via extremal perturbations and smooth masks. In *ICCV*, 2019.

[2] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu. Score-CAM: Score-weighted visual explanations for convolutional neural networks. In *CVPR Workshop*, 2020.