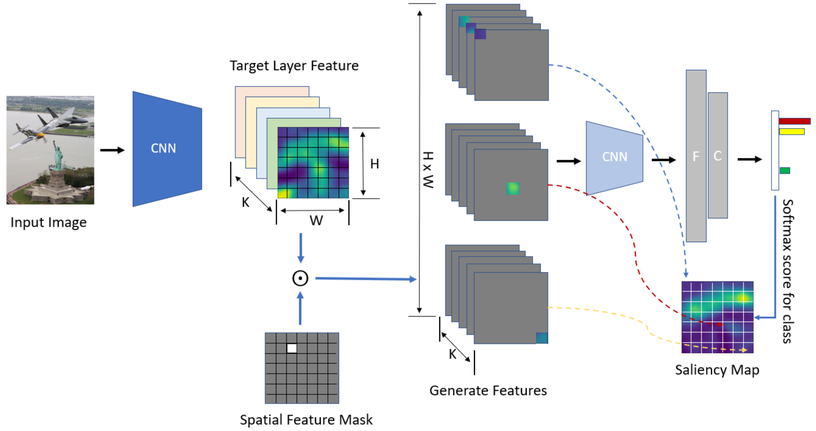
ReciproCAM: Lightweight Gradient-free Class Activation Map for Post-hoc Explanations

Seok-Yong Byun and Wonju Lee, Intel Corporation

Motivation

- There is a growing demand for interpretable or explainable AI (XAI), enabling the analysis of model behavior and the identification of potential bias or errors in a model or data.
- Class activation map (**CAM**) faced limitations in selecting appropriate model architectures.
- **GradCAM** and its variants were developed. However, these methods, relying on gradient calculations, encountered difficulties when applied in post-deployment frameworks.
- **ScoreCAM** shows state-of-the-arts localization performance, but it requires inferences equivalent to the number of feature maps. ScoreCAM is approximately 127 times slower than CAM.
- Meanwhile, black-box XAI algorithms like LIME, SHAP, or RISE have been proposed as post-hoc explanation schemes, which observes only input and output behaviors.
- **RISE** relies on Monte Carlo sampling of a random mask to approximate a true saliency map, typically requiring over thousands of inferences.
- **Extremal perturbation** introduces a gradient descent-based optimization method only for mask parameters. Nevertheless, this also consumes several seconds of GPU time.
- We here propose the lightweight gradient-free CAM method by employing reciprocal relationship between perturbed intermediate activations and the model output.

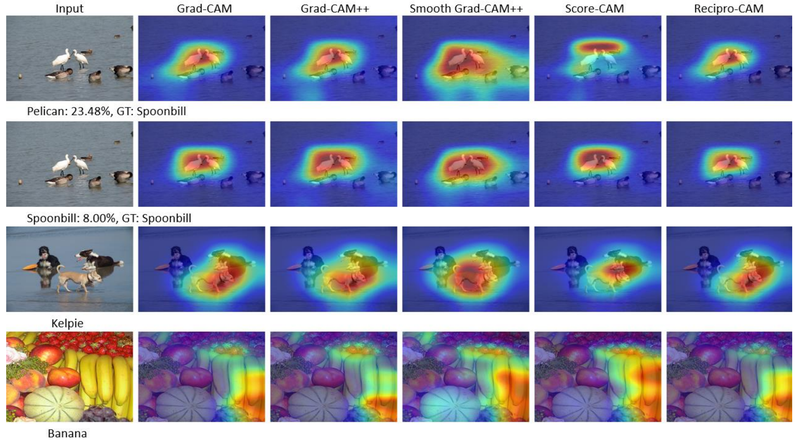
Method	Formula	Inference
CAM	$\mathrm{ReLU}\left(\sum_k w^{k,c}F^k\right)$	Single
GradCAM	$\mathrm{ReLU}\left(\sum_k \sum_{u,v} \frac{\partial y^c}{\partial f^k(u,v)} F^k\right)$	~ Double
ScoreCAM	$\mathrm{ReLU}\left(\sum_k h\left(F^k \odot x\right)^c F^k\right)$	~ $2K$ times
RISE	$\sum_n h(M_n \odot x)^c M_n$	N times



ReciproCAM

- **Spatially perturbed feature map generation**
  - Spatial mask M_n involves designating a single pixel in the feature map as 1, while setting all others to 0.
  - For a $H \times W$ feature map, total $N = H \times W$ spatial mask is generated.
  - Each spatial mask is uniquely associated with a specific pixel position in the feature map. The nth masked feature map corresponding to channel k is given by
    $$\tilde{F}_n^k = M_n \odot F^k.$$
  - By applying this explanatory hook in the middle of the network, we leverage not only the overlaps according to receptive field in the original input but also the reduced number of total mask N, resulting in faster execution and higher saliency resolution.
- **Saliency map generation**
  - By feedforwarding the nth spatially perturbed feature maps into a rest of network $g(\cdot)$, we obtain the logit for class c as
    $$y_n^c = g\left([\tilde{F}_n^1, \ldots, \tilde{F}_n^K]\right)^c.$$
  - We collect all N logits for class c, i.e., $\boldsymbol{y}^c = \left[y_1^c, \ldots, y_N^c\right]^T$, and perform the reshape operation to draw the saliency map as
    $$S^c = \mathrm{reshape}(\mathrm{norm}(\boldsymbol{y}^c), (H, W))$$

Experimental result



Method	Drop (↓)	Inc (↑)	Del (↓)	Ins (↑)	Coher (↑)	Compl (↓)	ADCC (↑)	Drop (↓)	Inc (↑)	Del (↓)	Ins (↑)	Coher (↑)	Compl (↓)	ADCC (↑)
	VGG-16							ResNet-18						
GradCAM	66.42	5.92	11.12	19.56	69.20	15.65	53.52	42.90	16.63	13.43	41.47	81.03	23.04	69.98
GradCAM++	32.88	20.10	8.82	36.60	89.34	26.33	75.65	17.85	34.46	12.30	44.80	98.18	44.63	74.24
SGradCAM++	36.72	16.11	10.57	31.36	82.68	28.09	71.72	20.67	29.99	12.83	43.13	97.53	43.11	74.20
ScoreCAM	26.13	24.75	9.52	47.00	93.83	20.27	**81.66**	12.81	40.41	10.76	46.01	98.35	41.78	77.30
ReciproCAM	21.51	34.86	9.50	46.88	92.24	27.48	80.27	20.68	36.30	10.19	44.93	97.38	33.60	**79.08**
	ResNet-50							ResNet-101						
GradCAM	32.99	24.27	17.49	48.48	82.80	22.24	75.27	29.38	29.35	18.66	47.47	81.97	22.51	76.40
GradCAM++	12.82	40.63	14.10	53.51	97.84	43.99	75.86	11.38	42.07	14.99	56.65	98.28	43.94	76.34
SGradCAM++	15.21	35.62	15.21	52.43	97.47	42.25	76.19	13.37	37.76	14.32	58.23	97.76	42.61	76.54
ScoreCAM	8.61	46.00	13.33	54.16	98.12	42.05	78.14	7.20	47.93	14.63	59.57	98.37	42.04	78.55
ReciproCAM	15.69	40.54	13.34	55.39	96.68	32.90	**80.84**	15.07	41.39	15.80	59.28	97.21	32.45	**81.38**
	ResNeXt-50							ResNeXt-101						
GradCAM	28.06	29.42	20.73	50.30	82.72	25.57	76.09	24.12	36.37	20.47	61.04	82.94	25.45	77.62
GradCAM++	11.12	41.38	17.07	56.06	97.30	48.66	73.16	9.74	42.63	17.63	62.90	95.05	46.27	74.61
SGradCAM++	12.70	36.58	16.90	56.76	97.32	47.48	73.58	9.49	40.43	17.67	64.16	96.81	49.24	73.03
ScoreCAM	7.20	45.70	15.59	57.92	98.00	46.86	75.38	5.37	47.70	17.30	63.61	97.03	46.83	75.60
ReciproCAM	13.70	40.82	18.94	58.93	96.37	37.36	**79.10**	12.03	42.69	20.25	64.70	97.50	35.62	**80.74**

- ReciproCAM achieves **state-of-the-arts results on ADCC**.
- ReciproCAM shows **1.2 and 147.8 times faster than GradCAM and ScoreCAM**, respectively.
- ReciproCAM is portable with the aid of explanatory hook for post deployment frameworks, e.g., ONNX or OpenVINO.
- Open source OpenVINO-XAI is coming soon!