SCAAT: Improving Neural Network Interpretability via Saliency Constrained Adaptive Adversarial Training

Rui Xu[1], Wenkang Qin[1], Peixiang Huang[1], Hao Wang[2], Lin Luo[1]*

[1] Peking University [2] National Institutes for Food and Drug Control, Beijing, China

BMVC 2023

Problem Definition and Contribution

Goal

Improving the deep learning model interpretability on the perspective of the quality of saliency predictions, while maintaining the model discriminative power.

Key Contributions

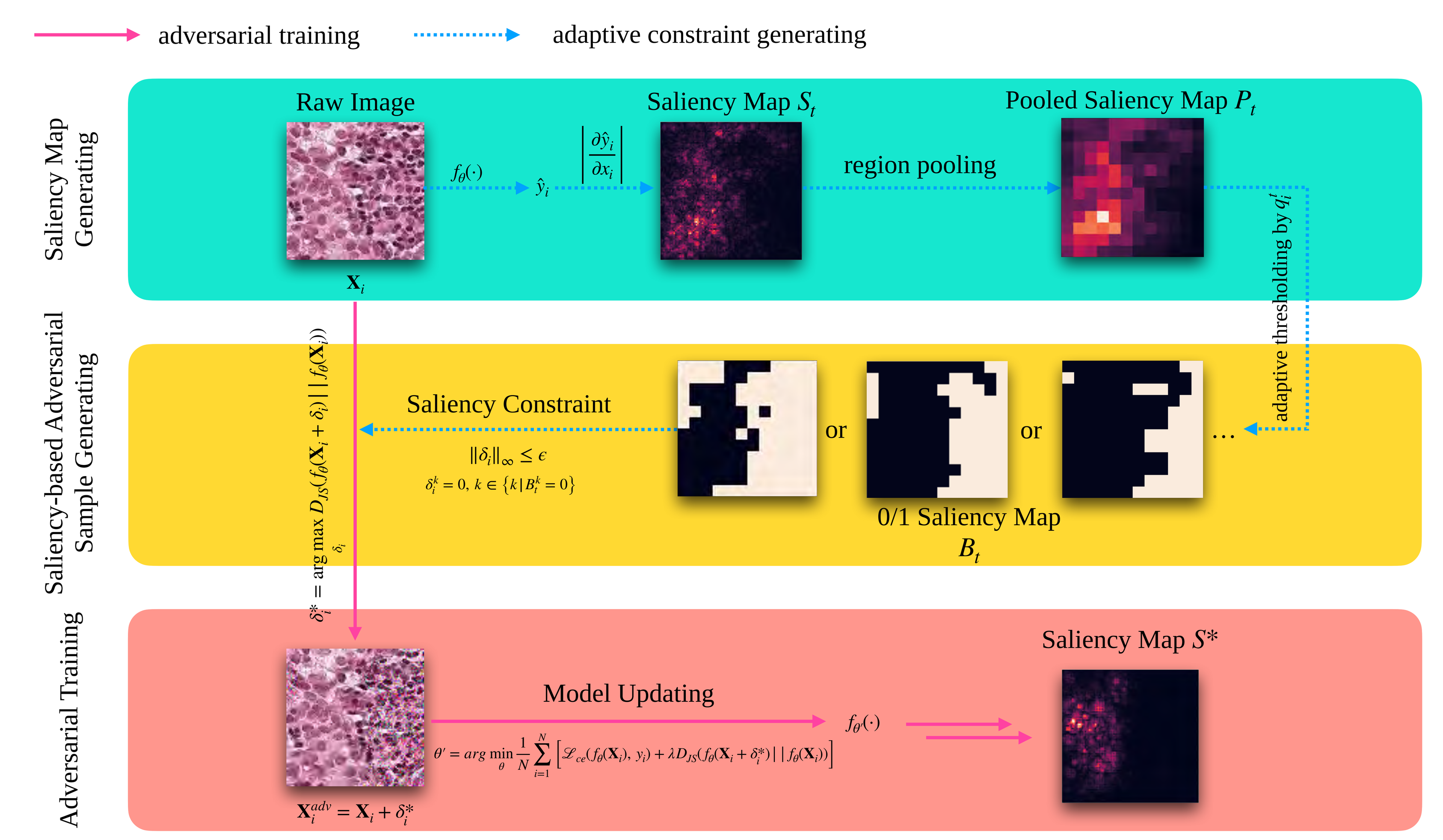
- A novel model-agnostic adaptive adversarial training framework which improves the interpretability of deep neural networks **without changing the network architectures**. The method can be generalized to various models and domains.
- Adaptive perturbation searching method which can **estimate the ratio of irrelevant features for each sample** then balance the learning performance and the quality of saliency predictions.
- SCAAT significantly improves the model interpretability in measures of saliency map sparsity and faithfulness on various natural and pathological image datasets, while **barely sacrificing the predictive performance of the models**.

Dataset

We conducted detailed experiments on CIFAR-10 and ImageNet-1k dataset for natural images, and PCAM dataset for pathological domain.

Method

Overall Pipeline of SCAAT

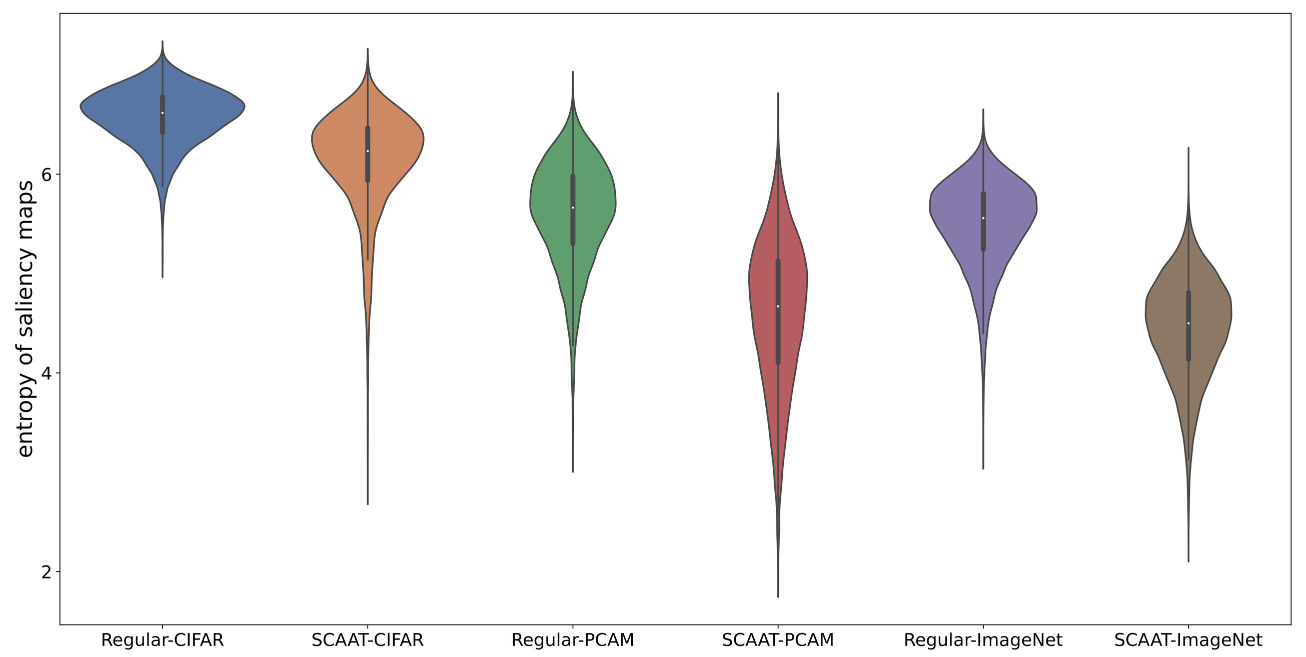


Irrelevant Feature Ratio Searching

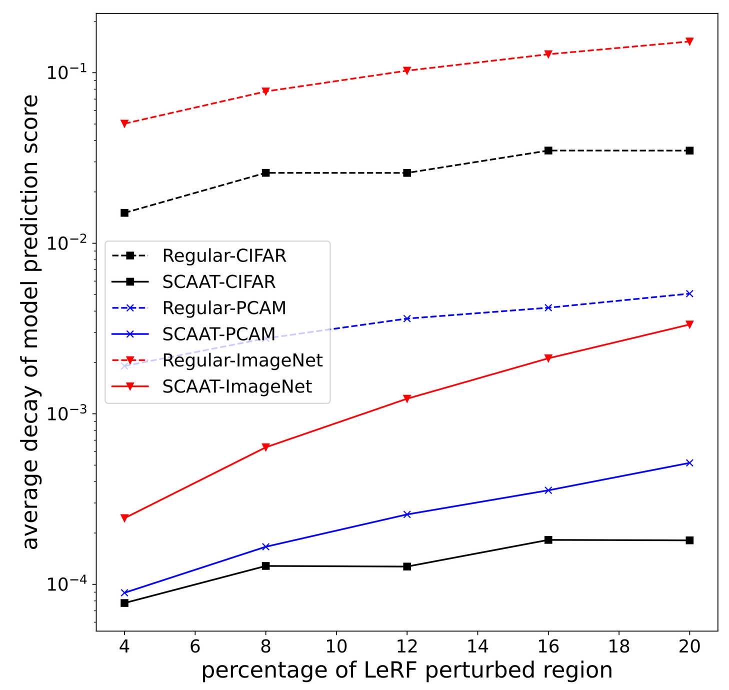
Rquire: i: Index of sample, $iter$: Iteration index; $N^{warm-up}$: Warm-up iterations;
Rquire: q_{max}, q_{min}: Boundary values for q; γ: Discretization for q searching.
if $iter \le N^{warm-up}$ **then**
  Set $q_i{}' = q_i$
else if $f_\theta(X_i^{adv})$ predicts as y_i **then**
  Set $q_i{}' = q_i + \gamma$
else
  Set $q_i{}' = q_i - \gamma$
end if
Set $q_i{}'' = \min(\max(q_i{}', q_{min}), q_{max})$
return $q_i{}''$

Experiments and Results

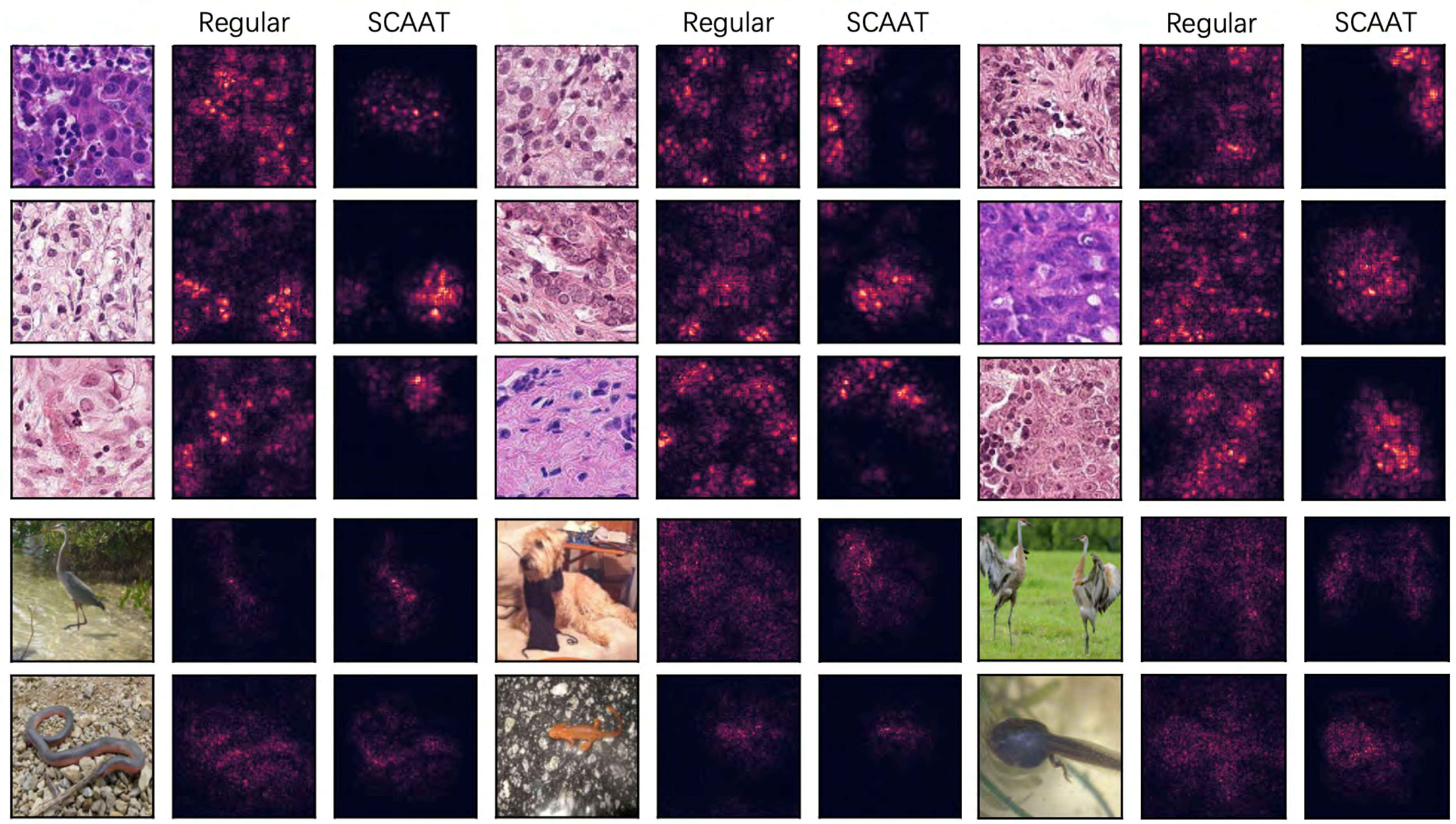
Saliency Map Entropy Distributions



Low-saliency Region Robustness Curve



Saliency Map Visualization on PCAM and ImageNet



Saliency Map Quality Comparison Across Datasets and Saliency Methods

	Evaluation Metric	Vallina Grad		Smooth Grad		Integrated Grad	
		Regular	SCAAT	Regular	SCAAT	Regular	SCAAT
PCAM	Saliency Entropy ↓	5.61	**4.56**	5.60	**4.54**	4.93	**4.43**
	Saliency Size (Kbyte) ↓	2.48	**1.61**	2.45	**1.61**	2.23	**1.52**
	$\text{AOPC}_{\texttt{lerf}}$↓ ($10^{-3}$)	3.20	**0.23**	2.89	**0.23**	8.94	**0.21**
	$\text{AOPC}_{\texttt{rel}}$ ↑	78.1	**1030**	90.0	**982**	24.6	**938**
ImageNet-1k	Saliency Entropy ↓	5.49	**4.45**	5.12	**4.23**	4.98	**4.15**
	Saliency Size (Kbyte) ↓	13.2	**7.12**	12.9	**6.94**	12.8	**6.80**
	$\text{AOPC}_{\texttt{lerf}}$↓ ($10^{-3}$)	85.2	**0.98**	72.5	**0.93**	43.2	**1.21**
	$\text{AOPC}_{\texttt{rel}}$ ↑	3.84	**321**	4.66	**346**	4.21	**305**