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

Rui Xu*[1]
xurui@stu.pku.edu.cn
Wenkang Qin*[1]
qinwk@stu.pku.edu.cn
Peixiang Huang[1,3]
huangpx@stu.pku.edu.cn
Hao Wang[2]
wanghao@nifdc.org.cn
Lin Luo†[1]
luol@pku.edu.cn

[1] College of Engineering
Peking University
Beijing, China

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

[3] Beijing Institute of Collaborative Innovation
Beijing, China

1 Appendix

1.1 Ablative Study

In this paper, we proposed an adversary-based training method to improve the model's saliency prediction via desensitizing it to irrelevant features in a self-supervised manner. The critical module of our work is about the adversarial sample construction from both the original image and its saliency map. Specifically, under the observation of that some images would have more irrelevant features than others, for each sample we adaptively adjust the proportion of features that are regarded as uncritical parts in the adversarial training process. Moreover, the lower bound $q_{\min}$ and upper bound $q_{\max}$ of the proportion q for samples are priorly assigned before training.

As $q_{\max}$ and $q_{\min}$ being empirically selected hyper-parameters, it should be careful for us to involve them into our framework. Thus we did detailed ablative experiments and finally demonstrated that the model is not very sensitive to them. From Table 1 we can learn that $q_{\min}$ behaves less critical to our SCAAT and can be set to 0.2 in most cases, while for $q_{\max}$ the values around 0.6 are usually acceptable and larger values may harm the model performance a lot.

Additionally, the adversarial training objective in our work is JS divergence rather than KL divergence or cross entropy as more common options. We also did ablative experiments for

*These authors contributed equally to this work.
†Corresponding author.

Table 1: **Ablative experiments of upper and lower bound of perturbation proportion q.** Experiments are done in the ImageNet-1k dataset and the model is ResNet-18.

Range of q	Top1 ACC (%)	Top5 ACC (%)	Sal. Entropy ↓	Sal. $\text{AOPC}_{\texttt{rel}}$ ↑
Traditional	68.70	88.35	5.489	3.843
[0.0, 0.4]	68.78	88.49	5.013	31.96
[0.0, 0.6]	68.41	88.27	4.764	109.6
[0.0, 0.8]	67.23	87.36	4.127	146.7
[0.2, 0.6]	68.21	87.98	4.448	321.2
[0.4, 0.6]	67.91	87.84	4.367	302.6

Table 2: **Ablative experiments of adversarial training objective.** Experiments are done in the ImageNet-1k dataset and the model is ResNet-18.

Training Objective	Top1 ACC (%)	Top5 ACC (%)	Sal. Entropy ↓	Sal. $\text{AOPC}_{\texttt{rel}}$ ↑
Cross Entropy	65.10	85.37	5.236	43.86
KL Divergence	67.33	87.15	4.679	189.4
JS Divergence	68.21	87.98	4.448	321.2

the training objective in Table 2. From the results we observed that JS divergence outperforms other metrics due to its symmetry and more stable gradients.

We also studied the effect of the range of perturbation (i.e. the ε-ball in adversarial training). As shown in Table 3, the ε of 0.02 leads to best performance with AUC of 0.930. And for the ε of 0.04 and 0.06, the $\text{AOPC}_{\texttt{rel}}$ increases largely with a slight drop to the AUC score. But ε greater than 0.10 will severely harm the model performance. The entropy of saliency maps keeps dropping when ε increases, but the $\text{AOPC}_{\texttt{rel}}$ is the comprehensive evaluation metric for the model interpretability, so we should mainly consider the ε with higher value of $\text{AOPC}_{\texttt{rel}}$. We observed that for different ε, there is a trade off between model performance and interpretability, and we finally set the ε to 0.08 as it makes the model much more interpretable and have comparable performance with the baseline.

1.2 Visualizations

Here in Figure 1 and Figure 2 we provide more visualization results to indicate that SCAAT improves the model's saliency prediction on both medical and natural image datasets.

Table 3: Performance of models trained against different scales of saliency constrained adversarial perturbation. Models are ResNet-18 and trained with the fixed $q = 0.5$.

Dataset	Setting of ε	AUC	Sal. Entropy	Sal. $\text{AOPC}_{\texttt{rel}}$
PCAM	$\varepsilon = 0.02$	0.930	5.47	282
	$\varepsilon = 0.04$	0.928	5.12	698
	$\varepsilon = 0.08$	0.925	4.93	974
	$\varepsilon = 0.10$	0.923	4.61	963
	$\varepsilon = 0.20$	0.895	4.38	731
	$\varepsilon = 0.40$	0.834	4.19	524

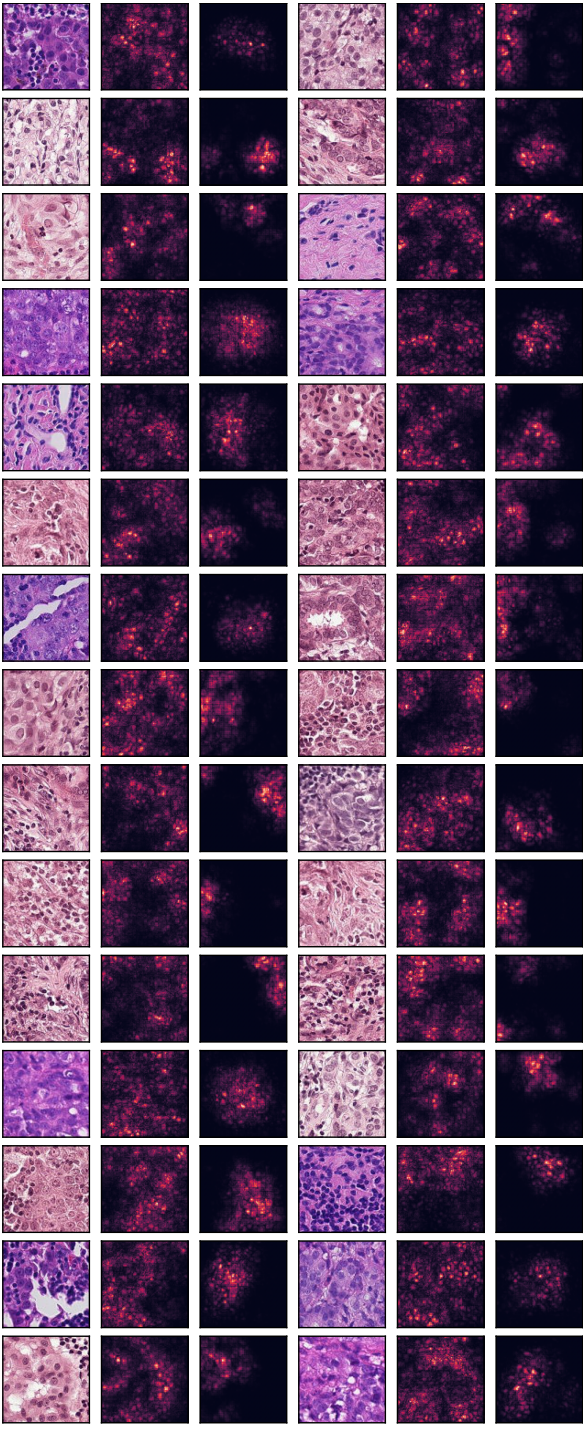

Figure 1: Visualizations of saliency maps of baseline model and ours on the PCAM dataset. Left pane is for the original image, middle pane is for baseline and the right pane is for our SCAAT.

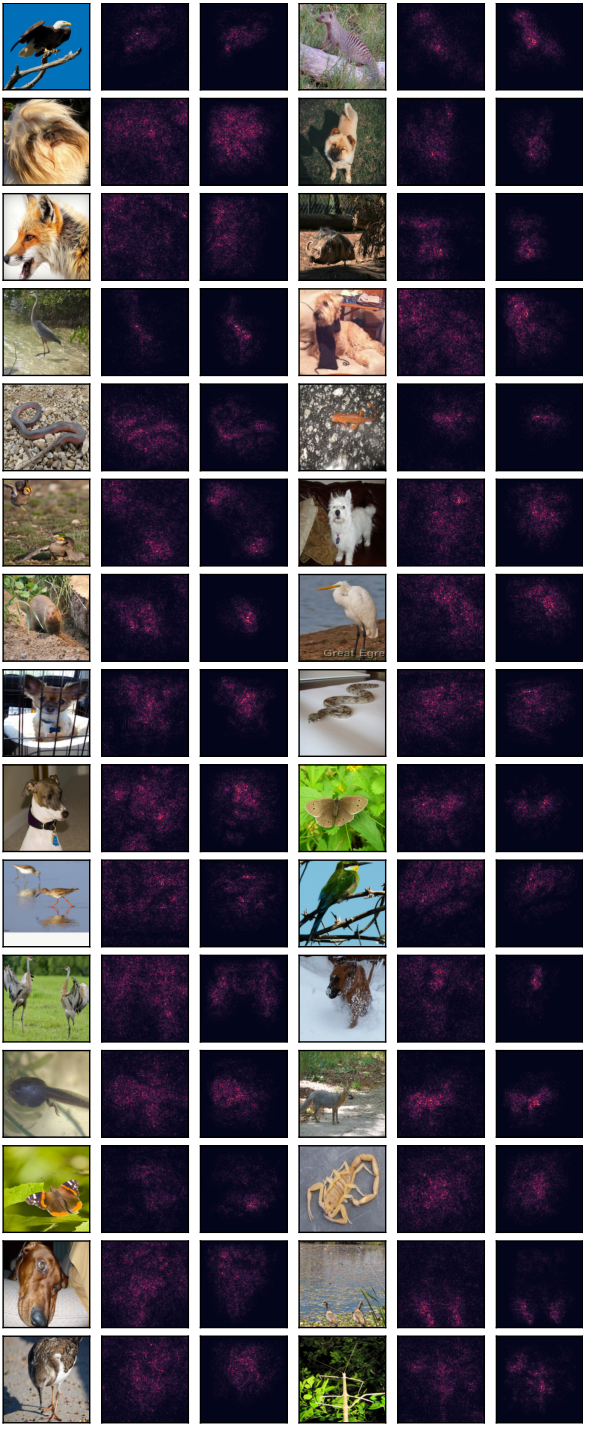

Figure 2: Visualizations on the ImageNet-1k dataset. Left pane is for the original image, middle pane is for baseline and the right pane is for our SCAAT.