Robust Attribution Regularization

Jiefeng Chen [1] Xi Wu [2] Vaibhav Rastogi [2] Yingyu Liang [1] Somesh Jha [1,3]

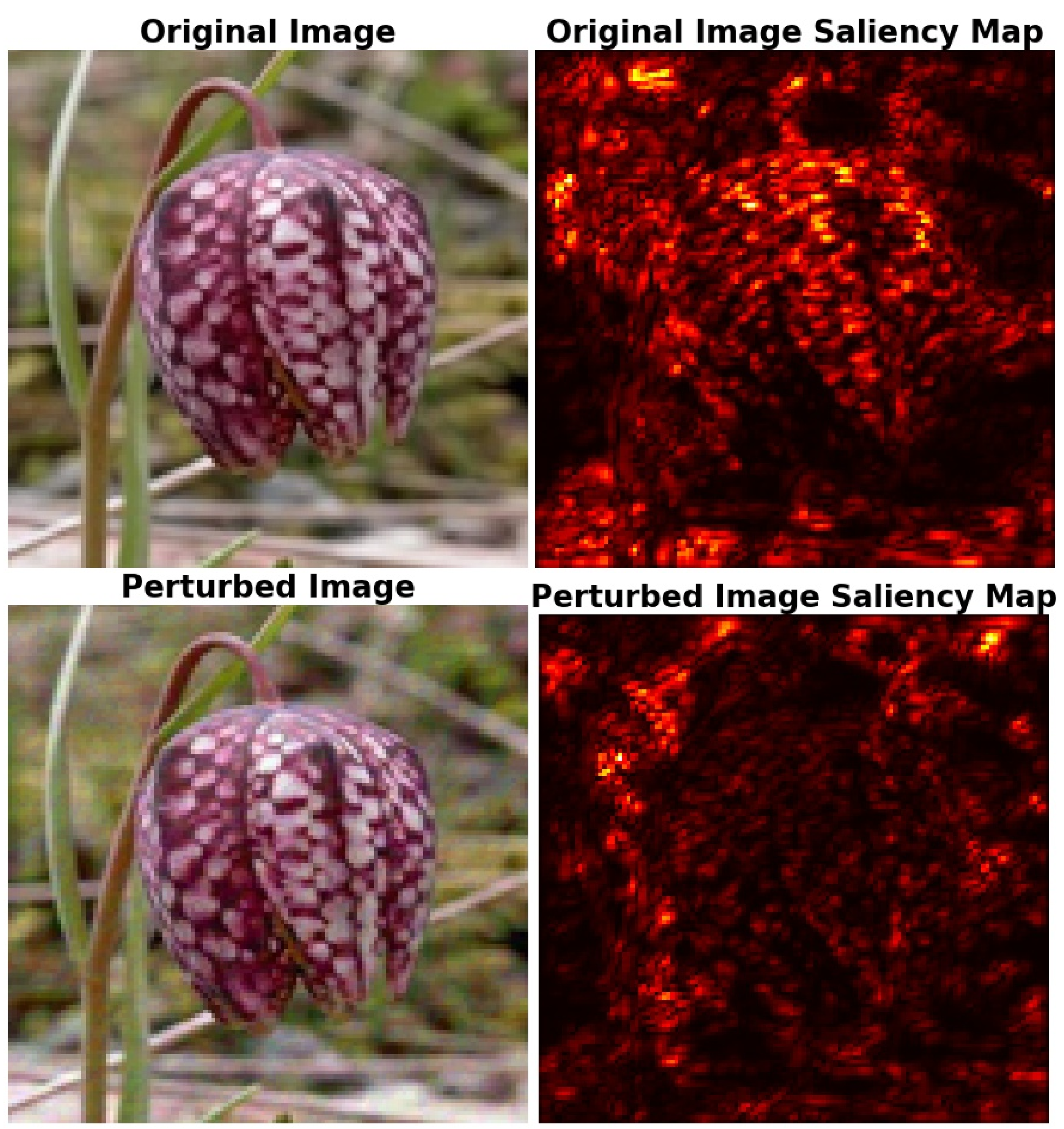
[1]University of Wisconsin-Madison [2]Google [3]XaiPient

Model Interpretations

An **attribution vector** indicates the importance of each feature in the input for the prediction. It can be computed via *Simple Gradient*, *DeepLIFT*, *Integrated Gradients(IG)*, etc.

Attribution of naturally trained model is brittle

Ghorbani et al. demonstrated that for existing models, one can generate **minimal perturbations** that **substantially change** model interpretations while **keeping their predictions intact**.



Top-1000 Intersection: 0.1%
Kendall's Correlation: 0.2607

Useful Information

Paper Link: https://arxiv.org/abs/1905.09957
Code link can be found in our paper!

RAR Training

We propose **Robust Attribution Regularization(RAR)** training to achieve robust attribution.

Uncertainty Set Model

$$\begin{aligned} &\underset{\theta}{\text{minimize}} \quad \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim P}[\rho(\mathbf{x},y;\theta)] \\ &\text{where } \ \rho(\mathbf{x},y;\theta) = \\ &\ell(\mathbf{x},y;\theta) + \lambda \max_{\mathbf{x}'\in N(\mathbf{x},\varepsilon)} s(\mathrm{IG}_{\boldsymbol{h}}^{\ell_y}(\mathbf{x},\mathbf{x}';r)) \end{aligned} \tag{1}$$

Refer to the paper for objectives in *Distributional Robustness Model*!

Instantiations

Classic Objectives are Weak Instantiations for Robust Attribution

- **Madry et al.'s Robust Prediction Objective:** Size function s() is sum(). Not a metric and allow attribution to cancel.
- **Input Gradient Regularization:** Only uses the first-term of IG for regularization.
- **Surrogate loss of Madry et al.'s min-max objective:** Regularizes by attribution of the loss output.

Strong Instantiations for Robust Attribution

- **IG-NORM:**

$$\min_{\theta} \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim P}[\ell(\mathbf{x},y;\theta) + \lambda \max_{\mathbf{x}'\in N(\mathbf{x},\varepsilon)} \|\mathrm{IG}^{\ell_y}(\mathbf{x},\mathbf{x}')\|_1]$$

- **IG-SUM-NORM:**

$$\min_{\theta} \mathop{\mathbb{E}}_{(\mathbf{x},y)\sim P}[\max_{\mathbf{x}'\in N(\mathbf{x},\varepsilon)} \ell(\mathbf{x}',y;\theta) + \beta\|\mathrm{IG}^{\ell_y}(\mathbf{x},\mathbf{x}')\|_1]$$

Read our paper to know how to set hyper-parameters to get these interesting instantiations!

1-Layer Neural Networks

Robust interpretation equals Robust prediction

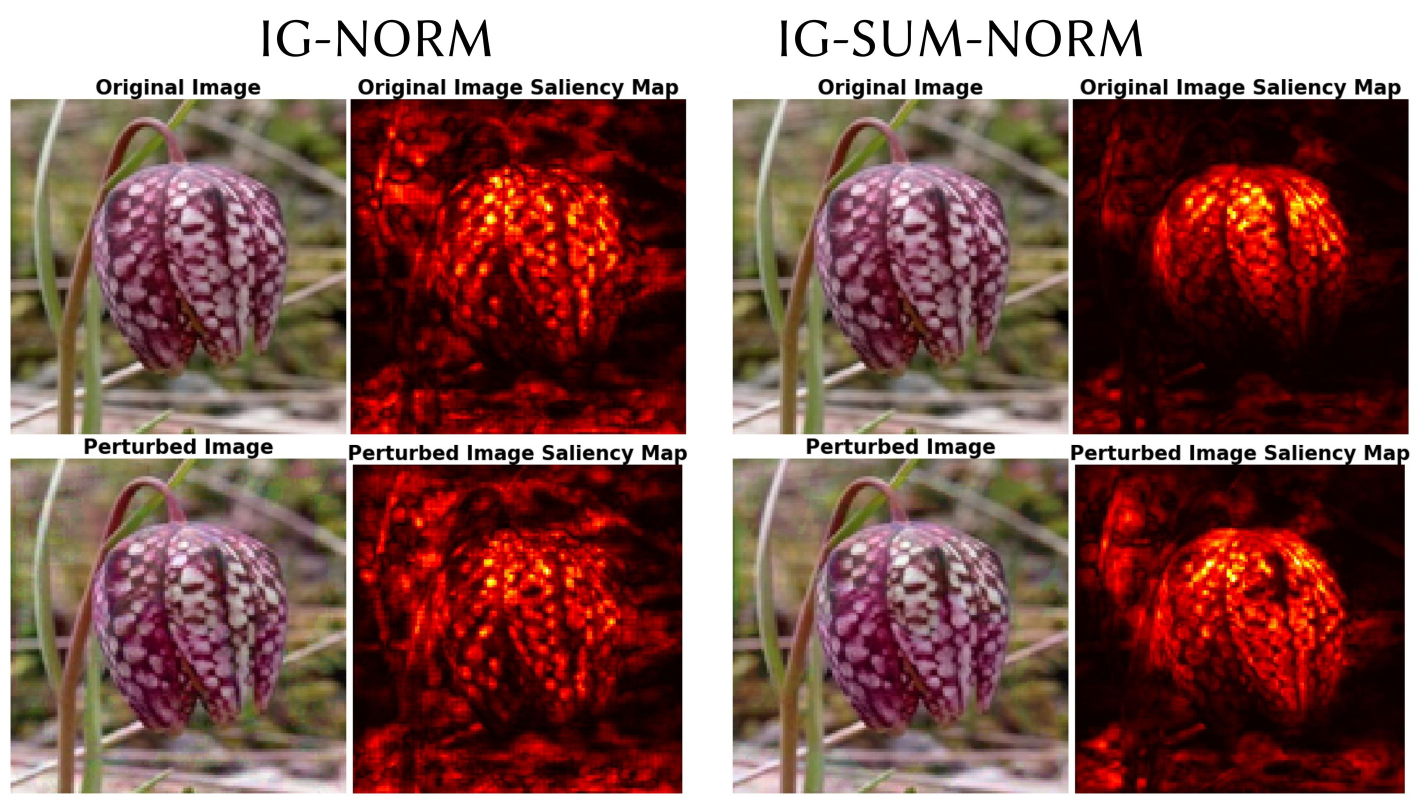
For the special case of one-layer neural networks, where the loss function takes the form of $\ell(\mathbf{x},y;\boldsymbol{w}) = g(-y\langle\boldsymbol{w},\mathbf{x}\rangle)$, the strong instantiations ($s(\cdot) = \|\cdot\|_1$) and weak instantiations ($s(\cdot) = sum(\cdot)$) coincide.

Read our paper for the details of our theories!

Empirical Results

Much more robust attribution using our technique!

Dataset	Approach	NA	AA	IN	CO
MNIST	NATURAL	99.17%	0.00%	46.61%	0.1758
	IG-NORM	98.74%	81.43%	71.36%	0.2841
	IG-SUM-NORM	98.34%	88.17%	**72.45%**	**0.3111**
GTSRB	NATURAL	98.57%	21.05%	54.16%	0.6790
	IG-NORM	97.02%	75.24%	**74.81%**	0.7555
	IG-SUM-NORM	95.68%	77.12%	74.04%	**0.7684**
Flower	NATURAL	86.76%	0.00%	8.12%	0.4978
	IG-NORM	85.29%	24.26%	64.68%	0.7591
	IG-SUM-NORM	82.35%	47.06%	**66.33%**	**0.7974**



IG-NORM: Top-1000 Intersection: 58.8%
Kendall's Correlation: 0.6736

IG-SUM-NORM: Top-1000 Intersection: 60.1%
Kendall's Correlation: 0.6951

More experimental results can be found in our paper!