Robust Attribution Regularization

Yingyu Liang

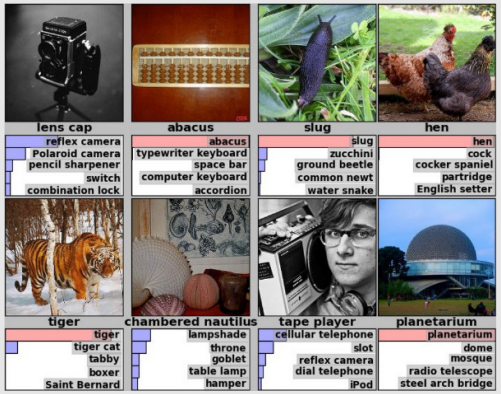
UW-Madison

Joint work with Jiefeng Chen, Xi Wu, Vaibhav Rastogi, and Somesh Jha

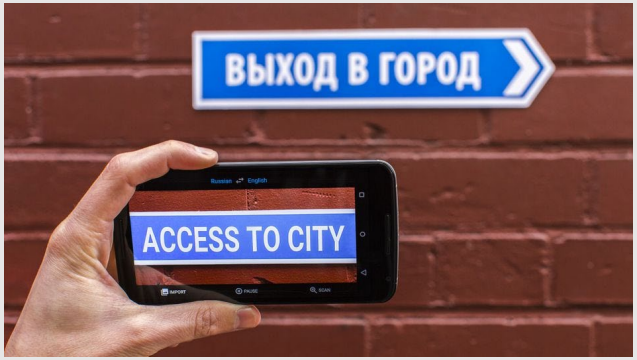
Appear in NeurIPS'2019

Machine Learning Progress

- Significant progress in Machine Learning

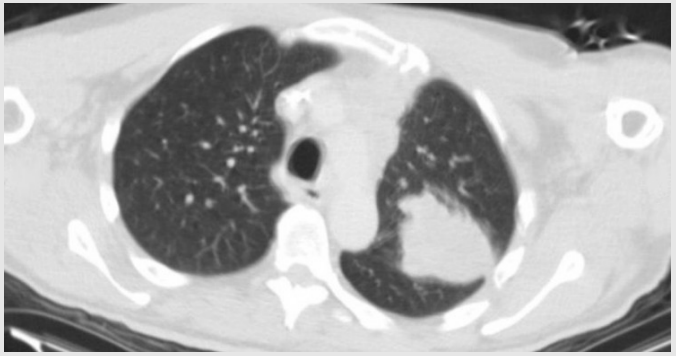


Computer vision



Machine translation

Game Playing

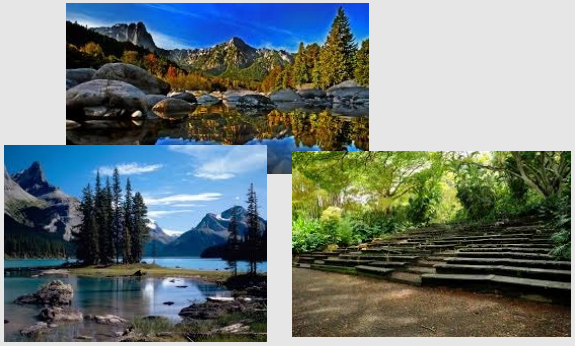


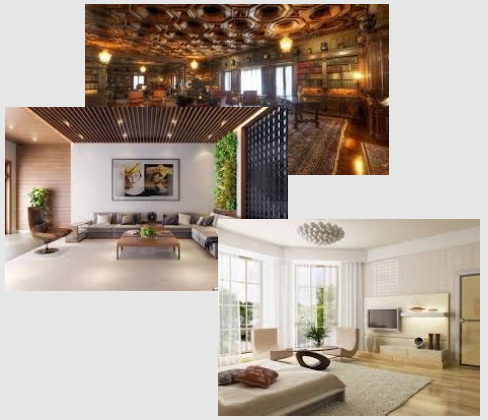
Medical Imaging

Key Engine Behind the Success

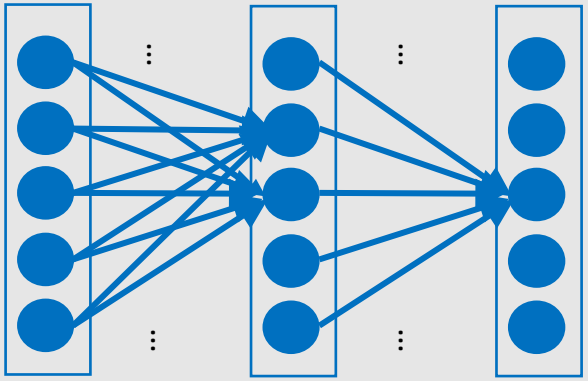
- Training Deep Neural Networks: $y = f(x; W)$
  - Given training data $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$
  - Try to find W such that the network fits the data

Outdoor

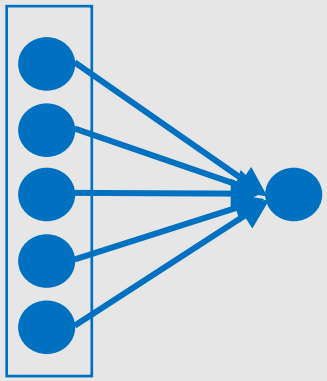




Indoor



... ...

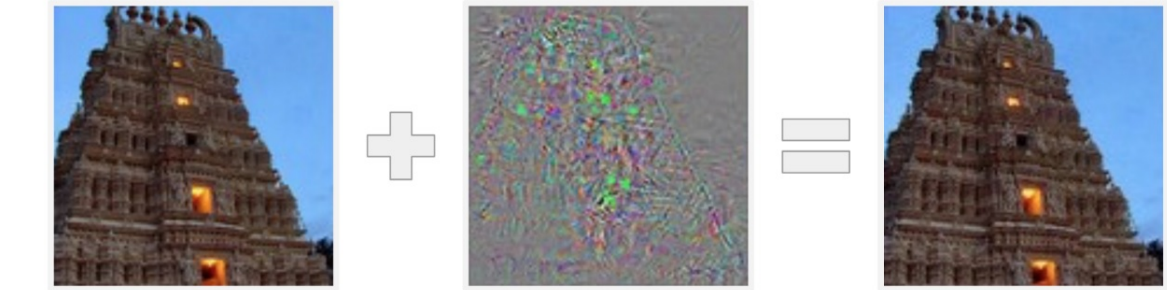


Outdoor

Challenges

- Blackbox: not too much understanding/interpretation

- Vulnerable to adversaries



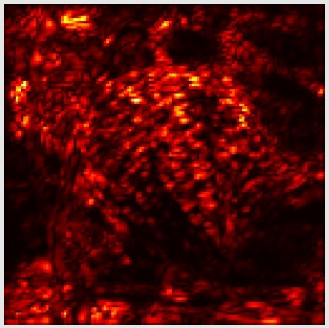
Interpretable Machine Learning

- Attribution task: Given a model and an input, compute an attribution map measuring the importance of different input dimensions

Machine Learning Model

Windflower

Compute Attribution



Integrated Gradient: Axiomatic Approach

Overview

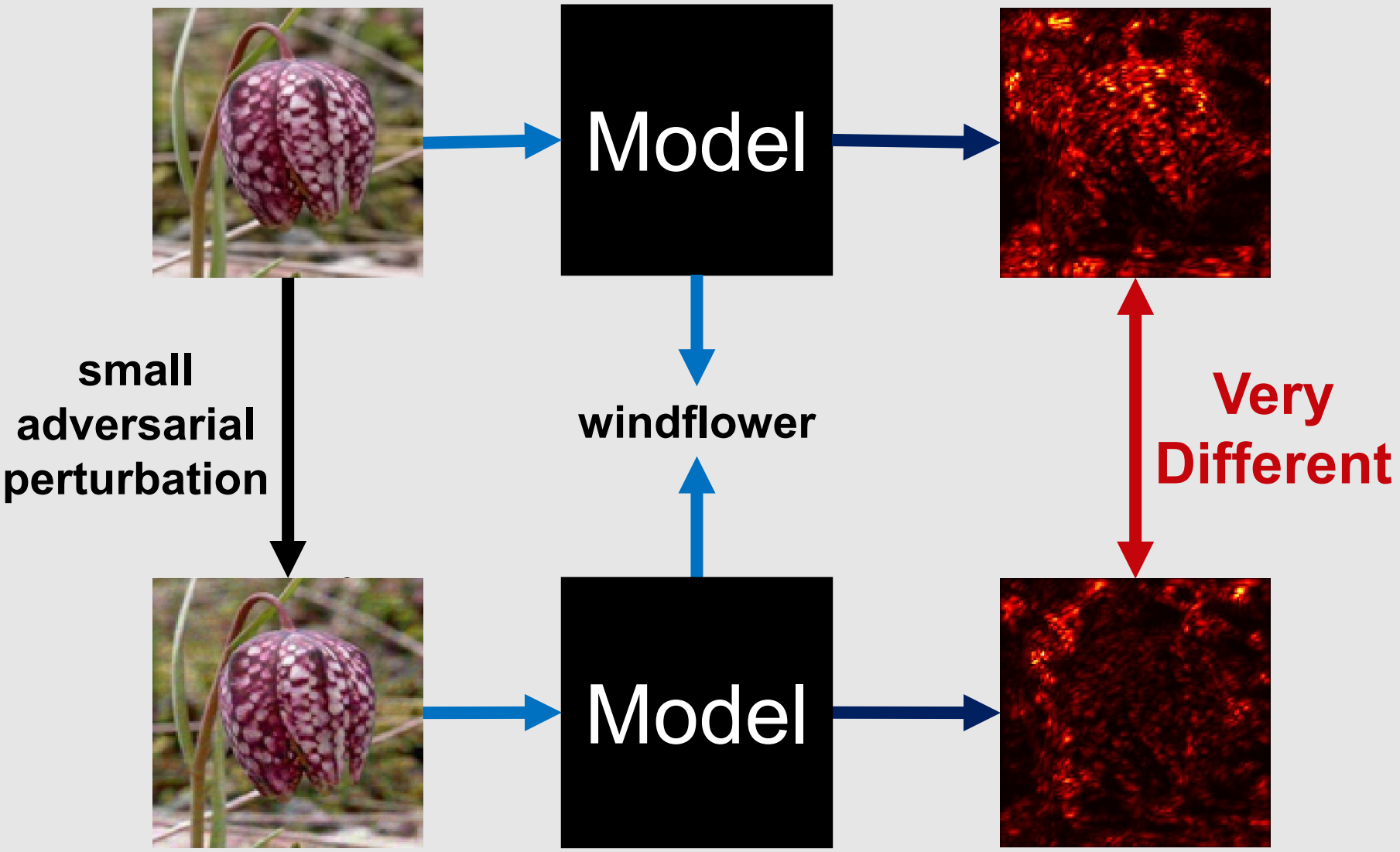
- List desirable criteria (axioms) for an attribution method
- Establish a uniqueness result: only this method satisfies these desirable criteria

- Inspired by economics literature: *Values of Non-Atomic Games*. Aumann and Shapley, 1974.

Axiomatic Attribution for Deep Networks.
Mukund Sundararajan, Ankur Taly, Qiqi Yan. ICML 2017.

Integrated Gradient: Example Results

Attribution is Fragile



Interpretation of Neural Networks is Fragile.
Amirata Ghorbani, Abubakar Abid, James Zou. AAAI 2019.

Robust Attribution Regularization

- Training for robust attribution: find a model that can get similar attributions for all perturbed images around the training image

$$\min_\theta \ \mathbb{E}[l(\boldsymbol{x}, y; \theta) + \lambda * \mathrm{RAR}]$$

$$\mathrm{RAR} = \max_{\boldsymbol{x}' \in \Delta(\boldsymbol{x})} s(\mathrm{IG}(\boldsymbol{x}, \boldsymbol{x}'))$$

Perturbed input

Allowed perturbations

Robust Attribution Regularization

- Training for robust attribution: find a model that can get similar attributions for all perturbed images around the training image

$$\min_\theta \ \mathbb{E}[l(\boldsymbol{x}, y; \theta) + \lambda * \mathrm{RAR}]$$

$$\mathrm{RAR} = \max_{\boldsymbol{x}' \in \Delta(\boldsymbol{x})} s(\mathrm{IG}(\boldsymbol{x}, \boldsymbol{x}'))$$

Size function

Integrated Gradient

Robust Attribution Regularization

- Training for robust attribution: find a model that can get similar attributions for all perturbed images around the training image

$$\min_\theta \ \mathbb{E}[l(\boldsymbol{x}, y; \theta) + \lambda * \mathrm{RAR}]$$

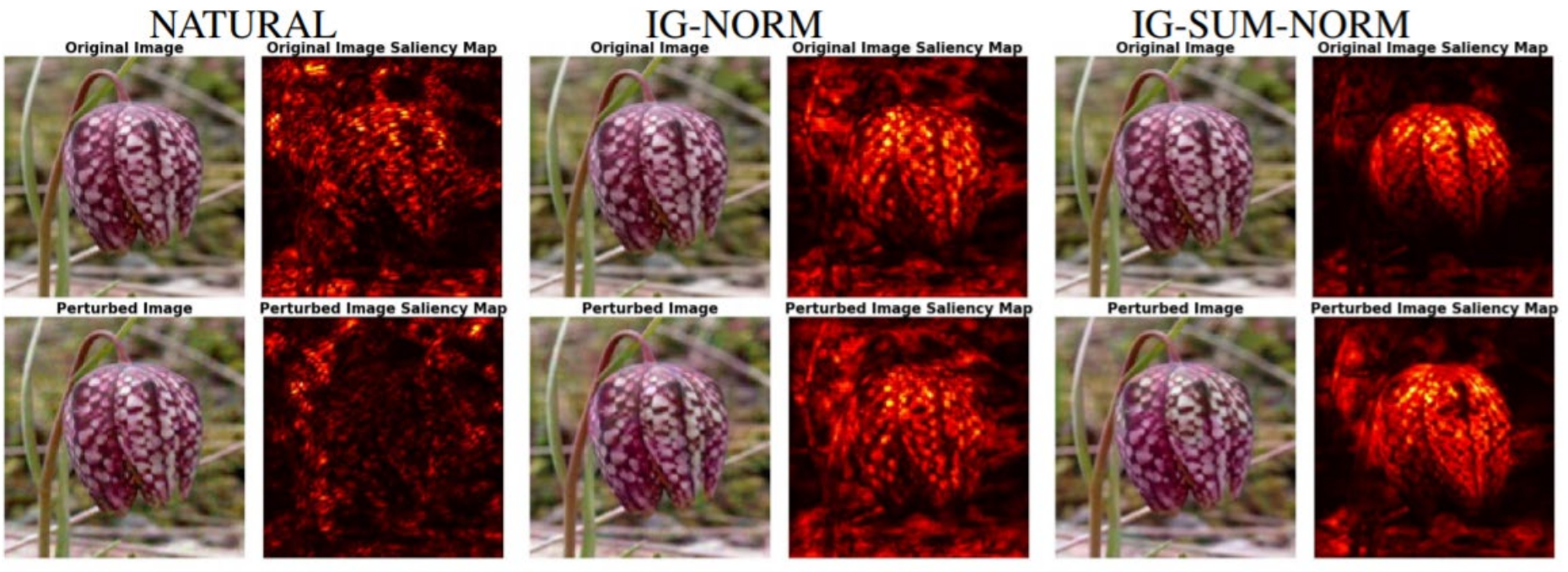
$$\mathrm{RAR} = \max_{\boldsymbol{x}' \in \Delta(\boldsymbol{x})} s(\mathrm{IG}(\boldsymbol{x}, \boldsymbol{x}'))$$

- Two instantiations:

$$\text{IG-NORM} = \max_{\boldsymbol{x}' \in \Delta(\boldsymbol{x})} \left|\left|\mathrm{IG}(\boldsymbol{x}, \boldsymbol{x}')\right|\right|_1$$

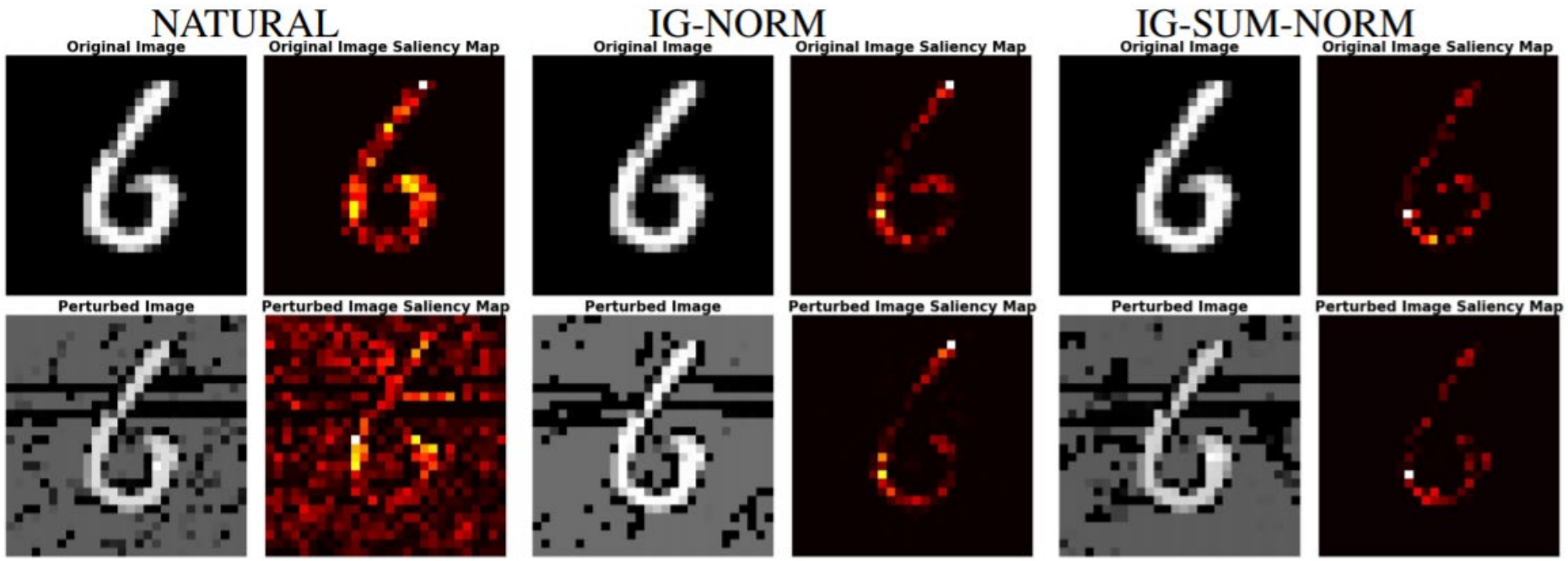
$$\text{IG-SUM-NORM} = \max_{\boldsymbol{x}' \in \Delta(\boldsymbol{x})} \left|\left|\mathrm{IG}(\boldsymbol{x}, \boldsymbol{x}')\right|\right|_1 + \mathrm{sum}(\mathrm{IG}(\boldsymbol{x}, \boldsymbol{x}'))$$

Experiments: Qualitative



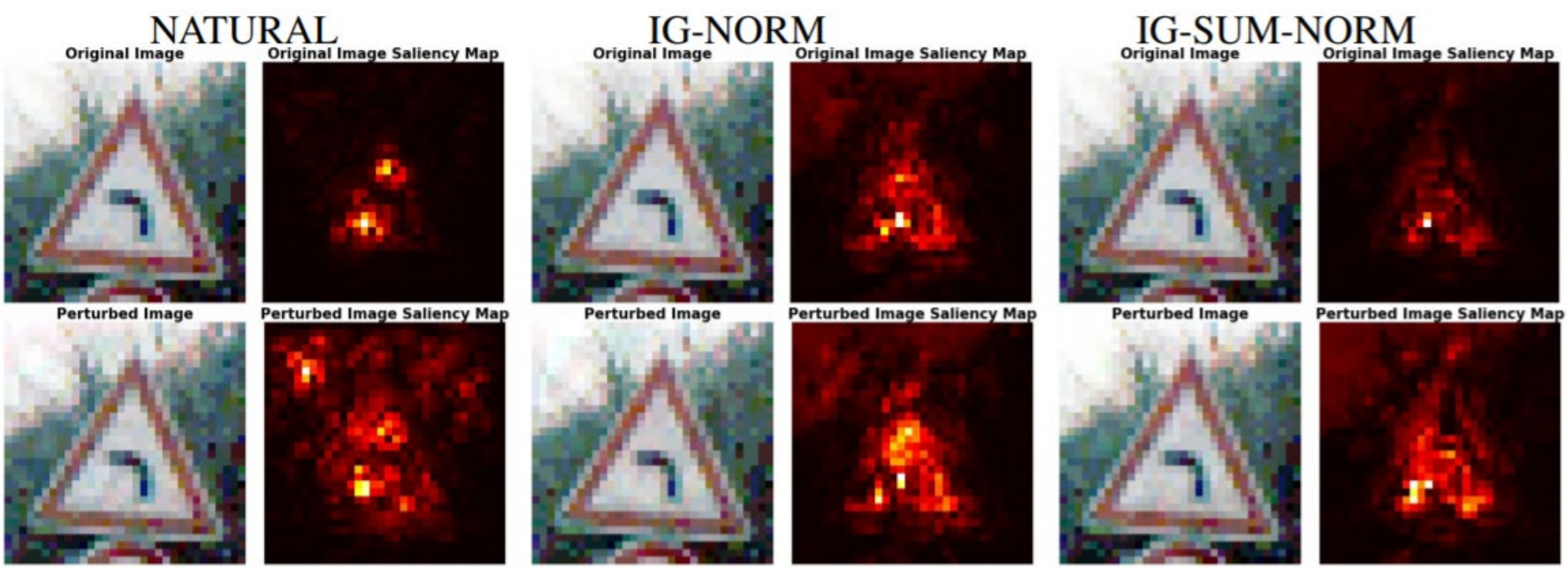
Flower dataset

Experiments: Qualitative



MNIST dataset

Experiments: Qualitative

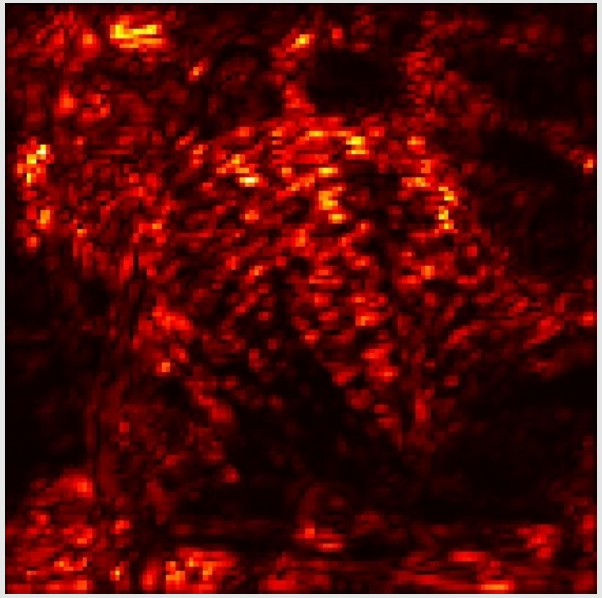


GTSRB dataset

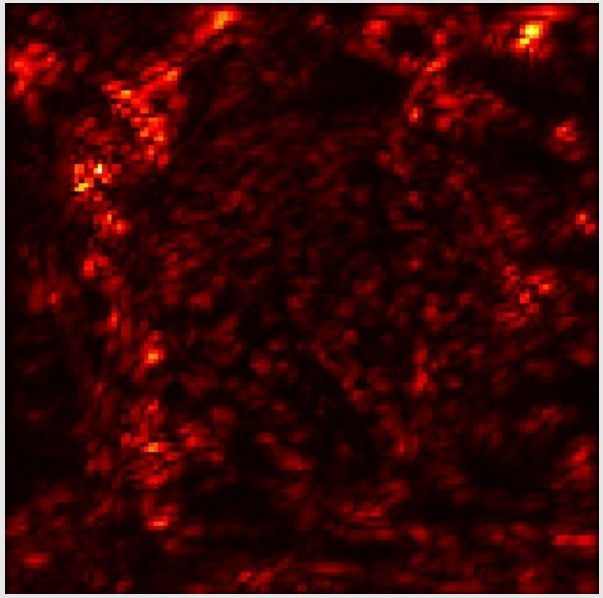
Experiments: Quantitative

- Metrics for attribution robustness
  1. Kendall's tau rank order correlation
  2. Top-K intersection

Original Image Attribution Map

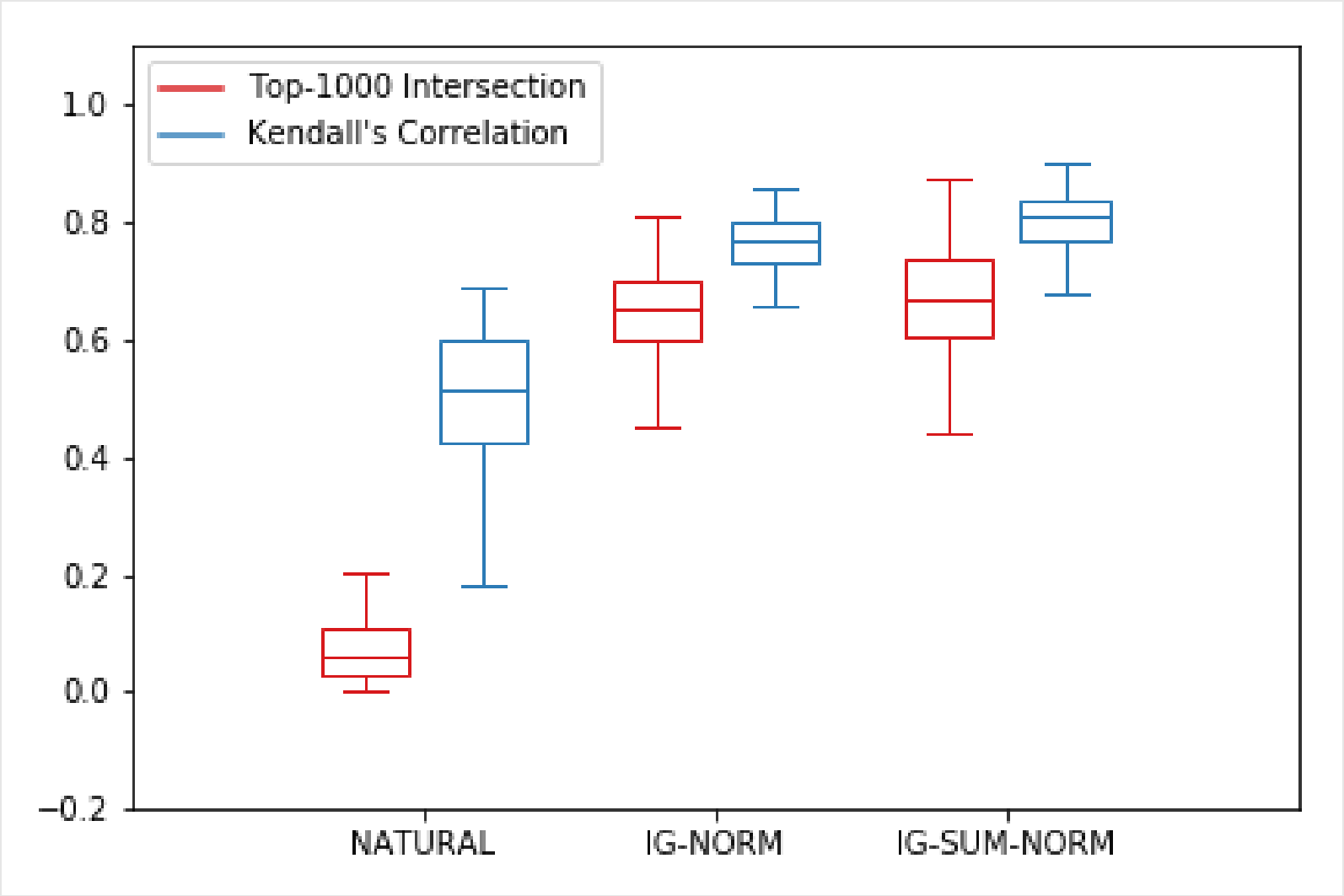


Perturbed Image Attribution Map

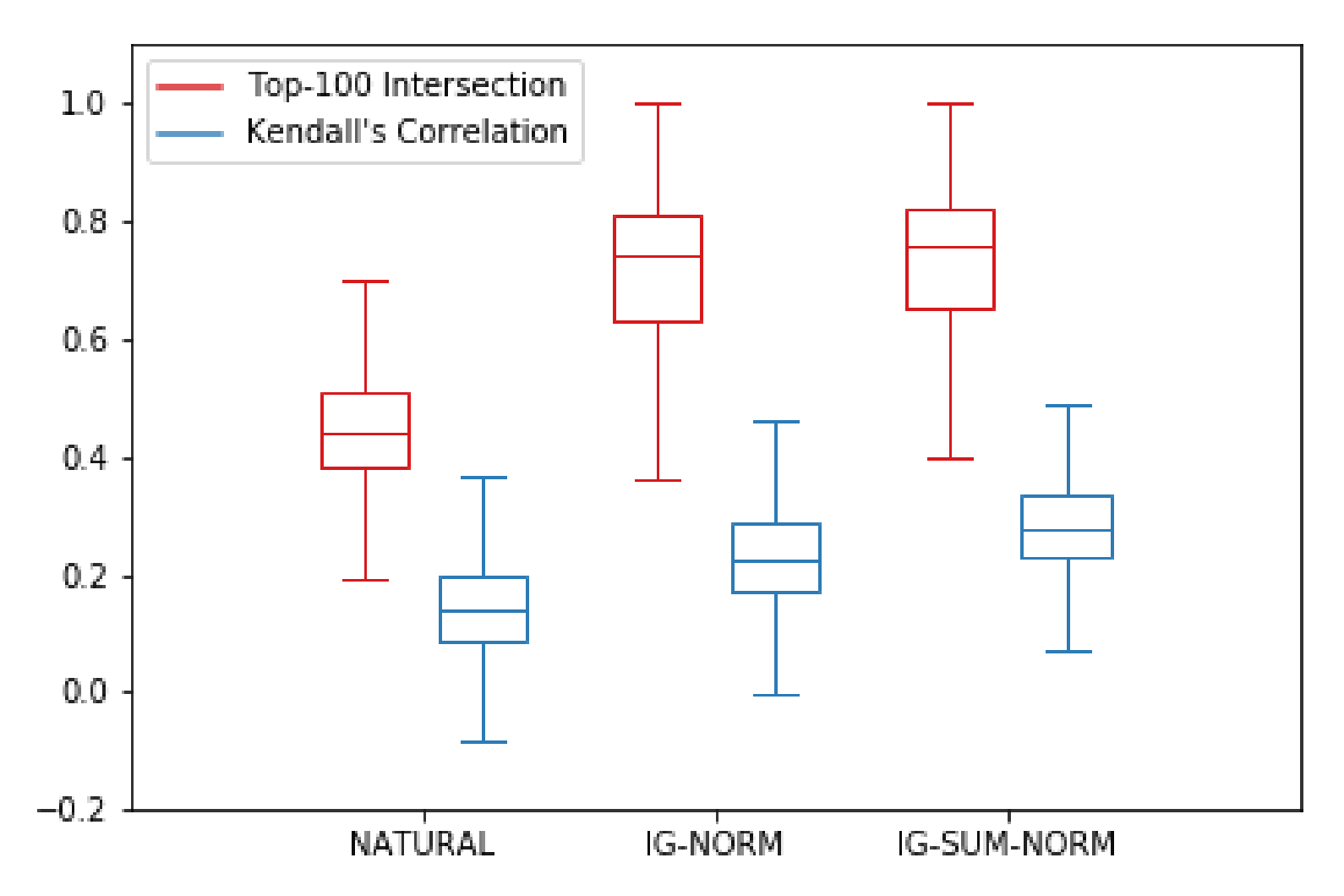


Top-1000 Intersection: 0.1%
Kendall's Correlation: 0.2607

Result on Flower dataset



Result on MINST dataset



Result on GTSRB dataset