Robust Attribution Regularization

Yingyu Liang

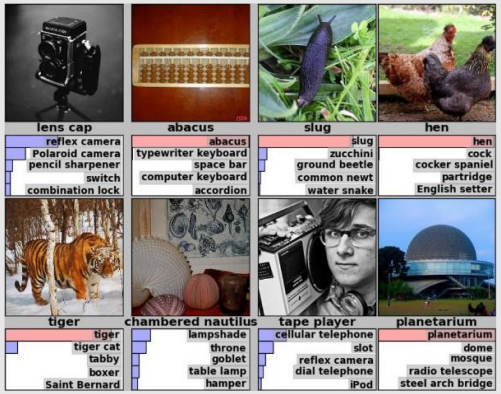
UW-Madison

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

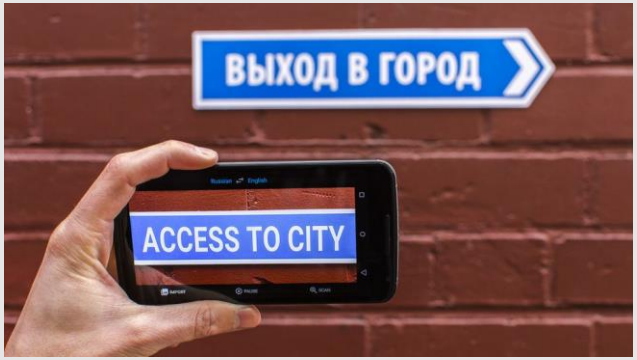
Appear in NeurIPS'2019

Machine Learning Progress

- Significant progress in Machine Learning

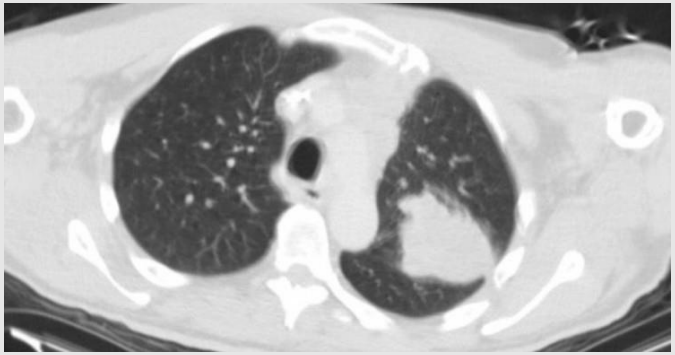


Computer vision



Machine translation

Game Playing

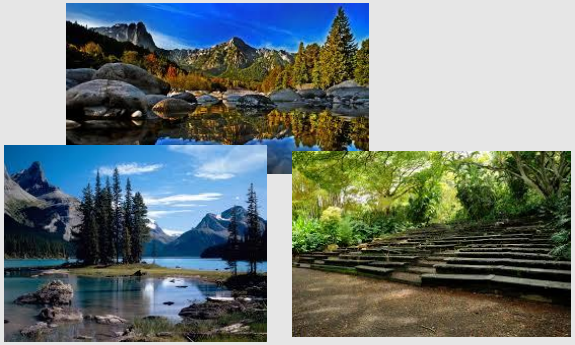


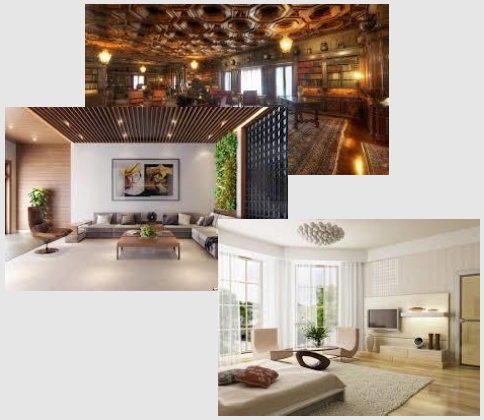
Medical Imaging

Key Engine Behind the Success

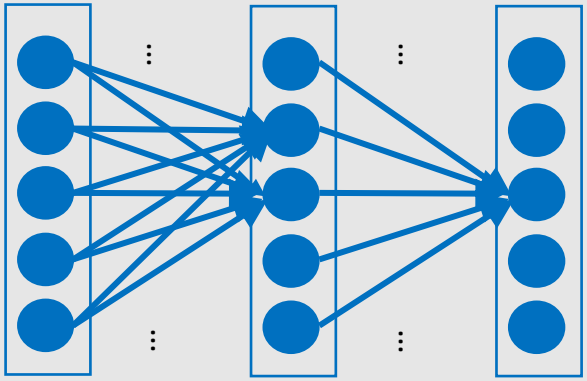
- Training Deep Neural Networks: $y = f(x; W)$
  - Given training data $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$
  - Try to find W such that the network fits the data

Outdoor

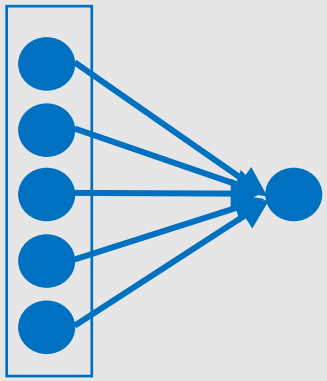




Indoor



… …



Outdoor

Challenges

- Blackbox: not too much understanding/interpretation

Black Box

Windflower

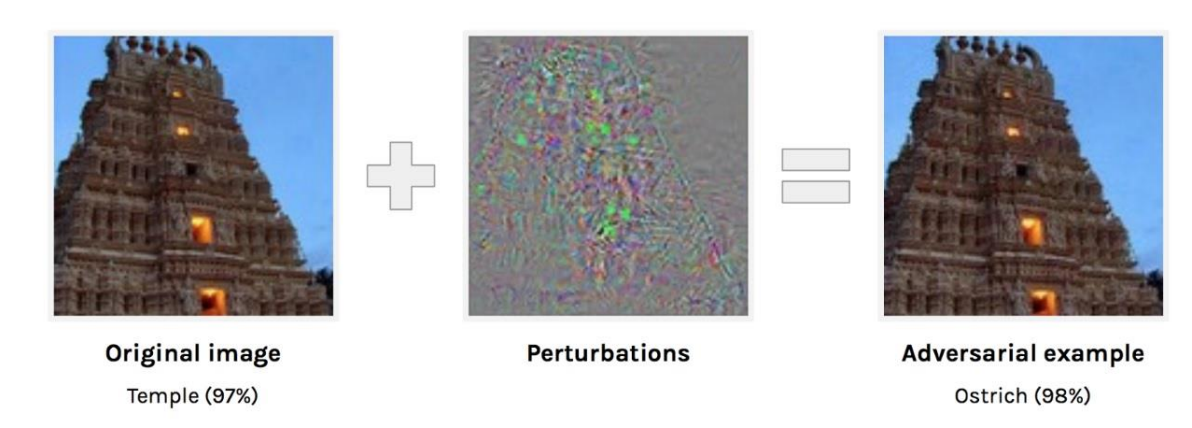
Challenges

- Blackbox: not too much understanding/interpretation

Black Box

Windflower

- Vulnerable to adversaries



Interpretable Machine Learning

- Attribution task: Given a model and an input

Machine Learning Model

Windflower

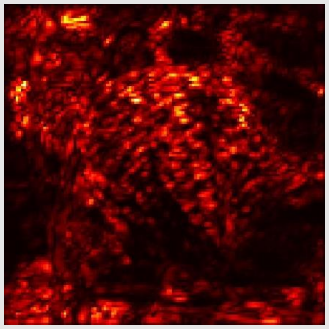
Interpretable Machine Learning

- 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: Axioms

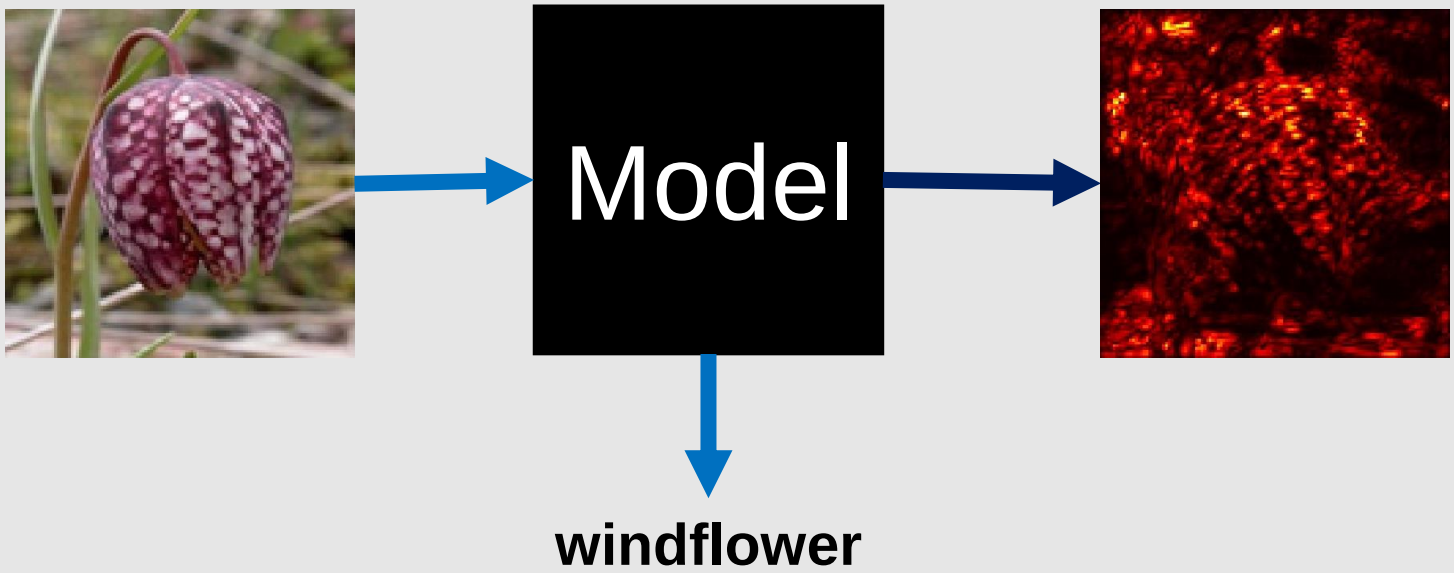
- **Implementation Invariance:** Two networks that compute identical functions for all inputs get identical attributions even if their architecture/parameters differ

- **Sensitivity:**
- (a) If baseline and input have different scores, but differ in a single variable, then that variable gets some attribution
- (b) If a variable has no influence on a function, then it gets no attribution

- **Linearity preservation:** Attr(a*f1 + b*f2)=a*Attr(f1)+b*Attr(f2)
- **Completeness:** sum(Attr) = f(input) – f(baseline)
- **Symmetry Preservation:** Symmetric variables with identical values get equal attributions

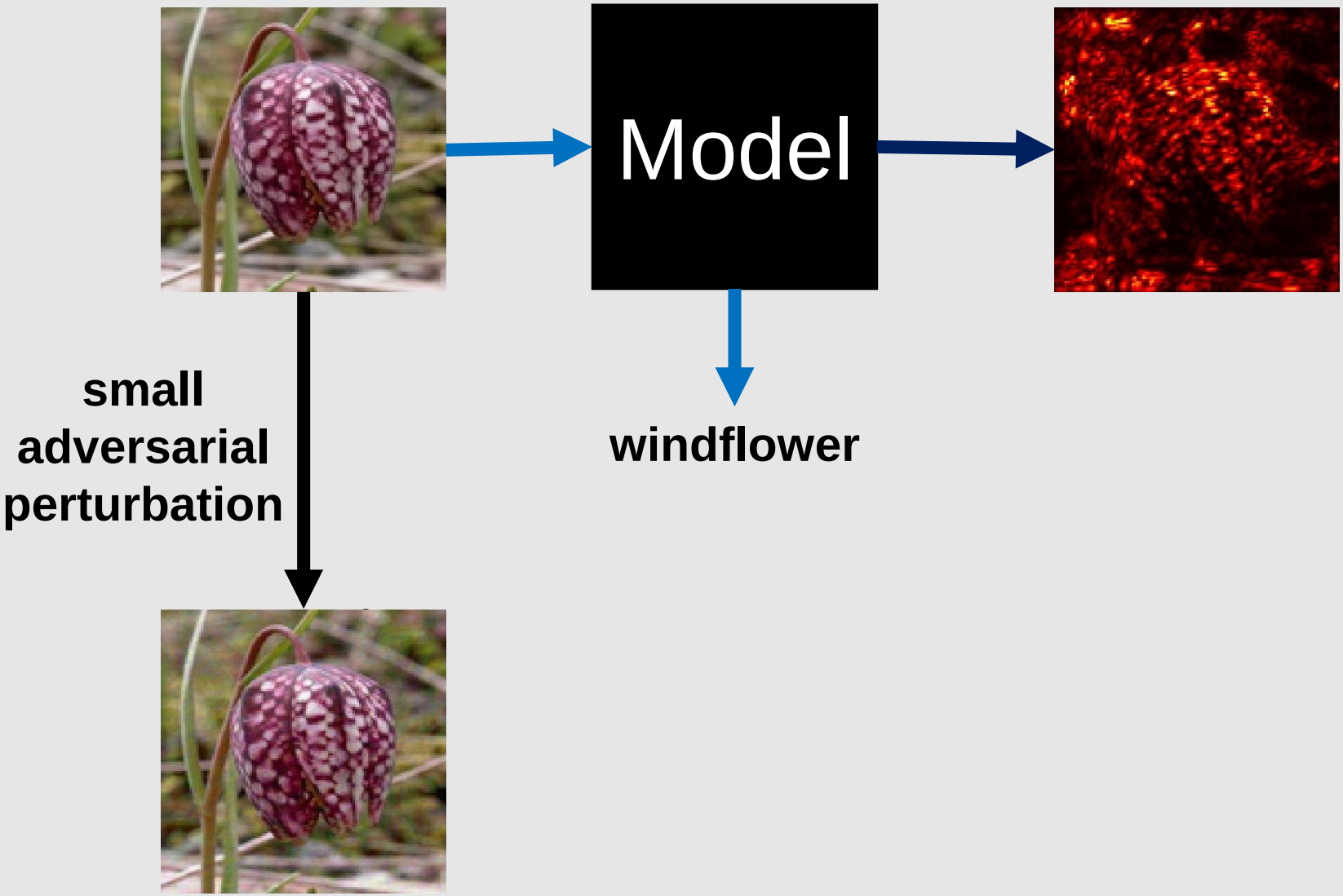
Integrated Gradient: Example Results

Observation 1: Attribution is Fragile



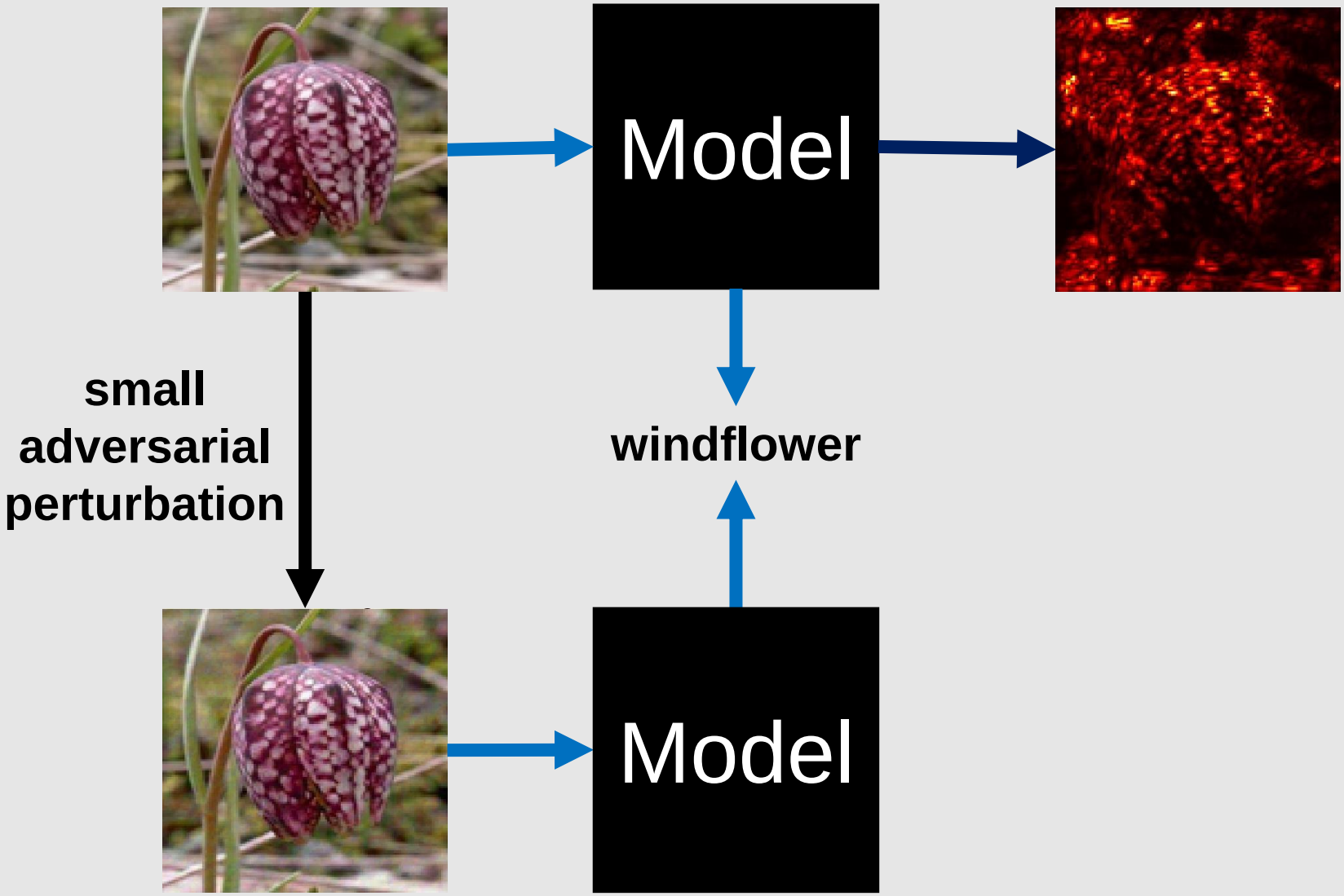
Interpretation of Neural Networks is Fragile.
Amirata Ghorbani, Abubakar Abid, James Zou. AAAI 2019.

Observation 1: Attribution is Fragile



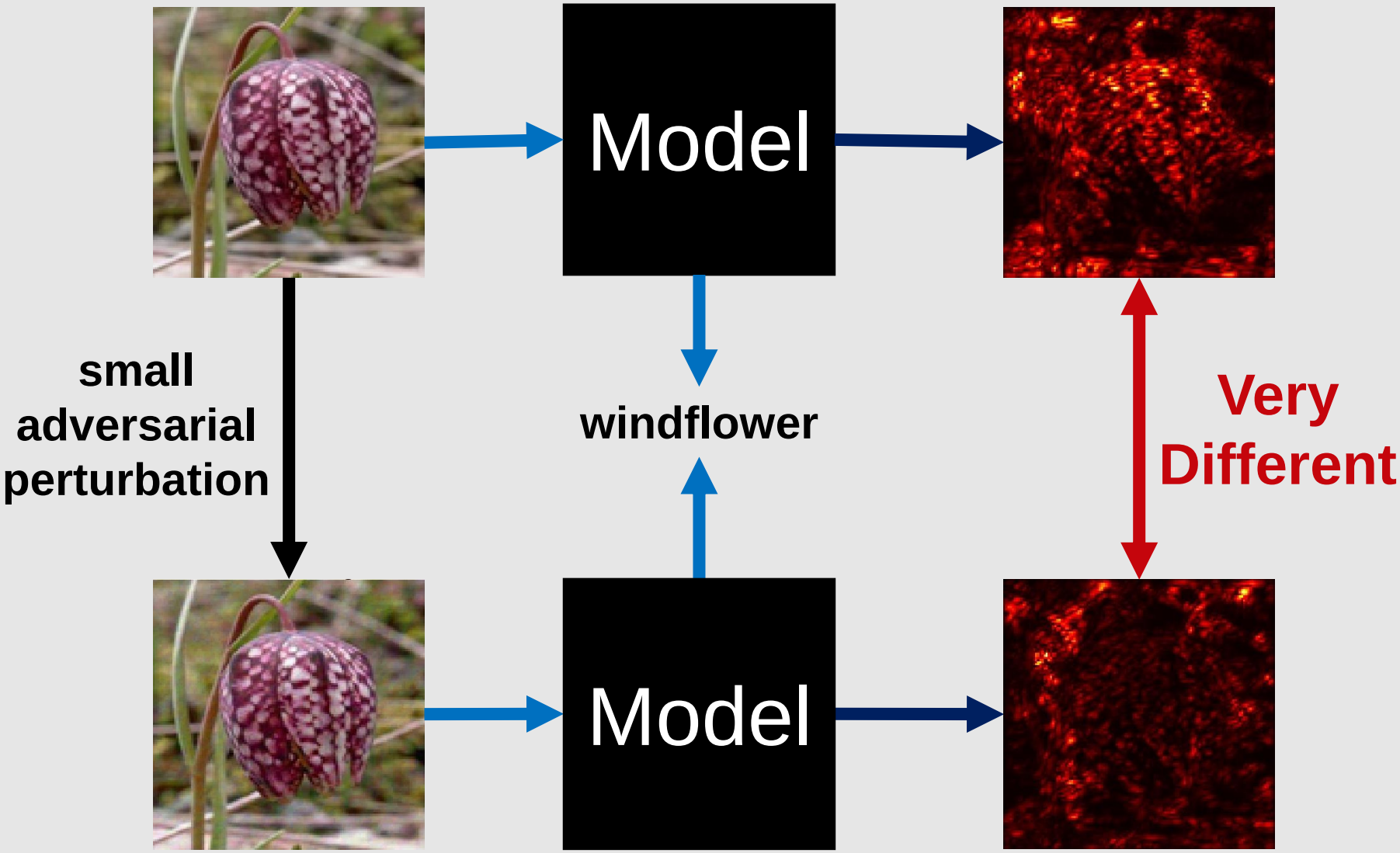
Interpretation of Neural Networks is Fragile.
Amirata Ghorbani, Abubakar Abid, James Zou. AAAI 2019.

Observation 1: Attribution is Fragile



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

Observation 1: Attribution is Fragile

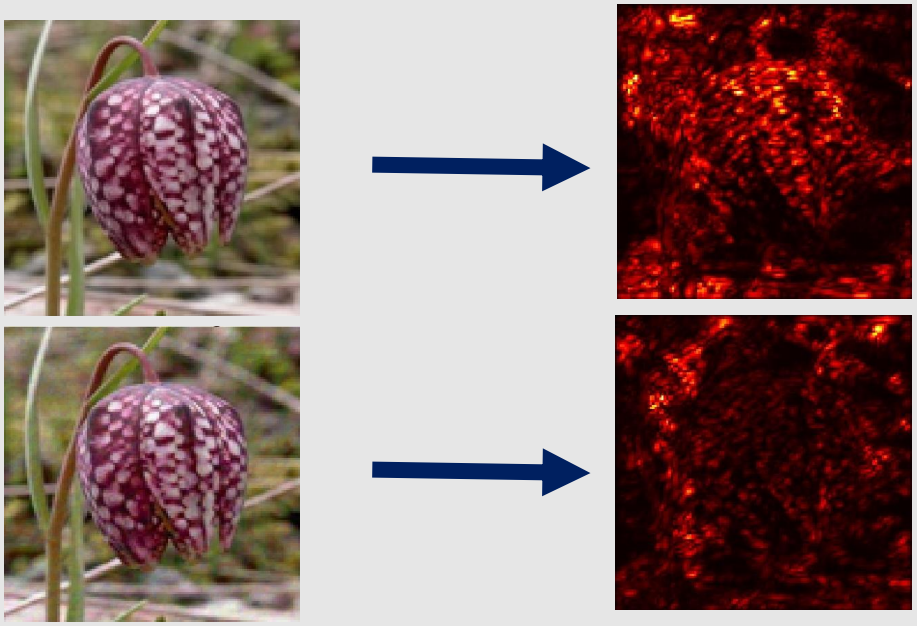


Interpretation of Neural Networks is Fragile.
Amirata Ghorbani, Abubakar Abid, James Zou. AAAI 2019.

Observation 2: Robust Prediction Correlates with Robust Attribution

- Normally trained model

original image,
normally trained model



perturbed image,
normally trained model

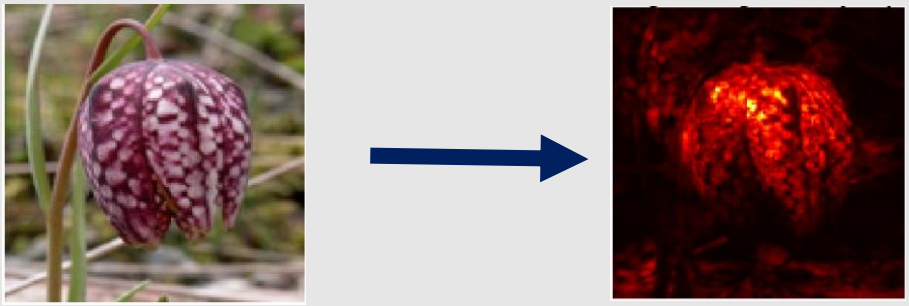
Observation 2: Robust Prediction Correlates with Robust Attribution

- Training for robust prediction: find a model that predicts the same label for all perturbed images around the training image

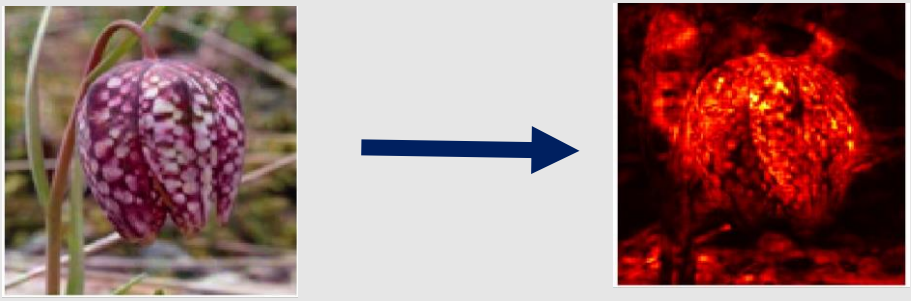
Observation 2: Robust Prediction Correlates with Robust Attribution

- Training for robust prediction: find a model that predicts the same label for all perturbed images around the training image

original image,
robustly trained model



perturbed image,
robustly trained model



Robust Attribution Regularization

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

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

Robust Attribution Regularization

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

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

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

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

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

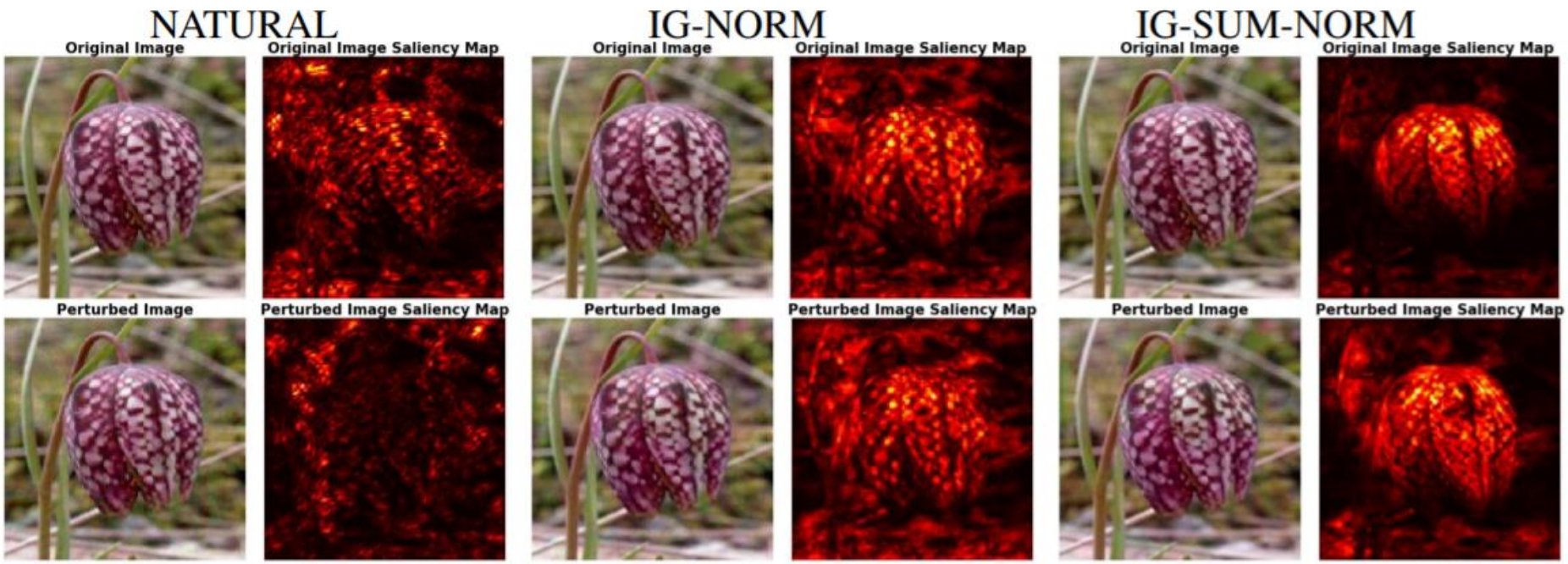
$$\mathrm{RAR} = \max_{\boldsymbol{x}' \in \Delta(\boldsymbol{x})} s(\mathrm{IG}(\boldsymbol{x}, \boldsymbol{x}'))$$

- Two instantiations:

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

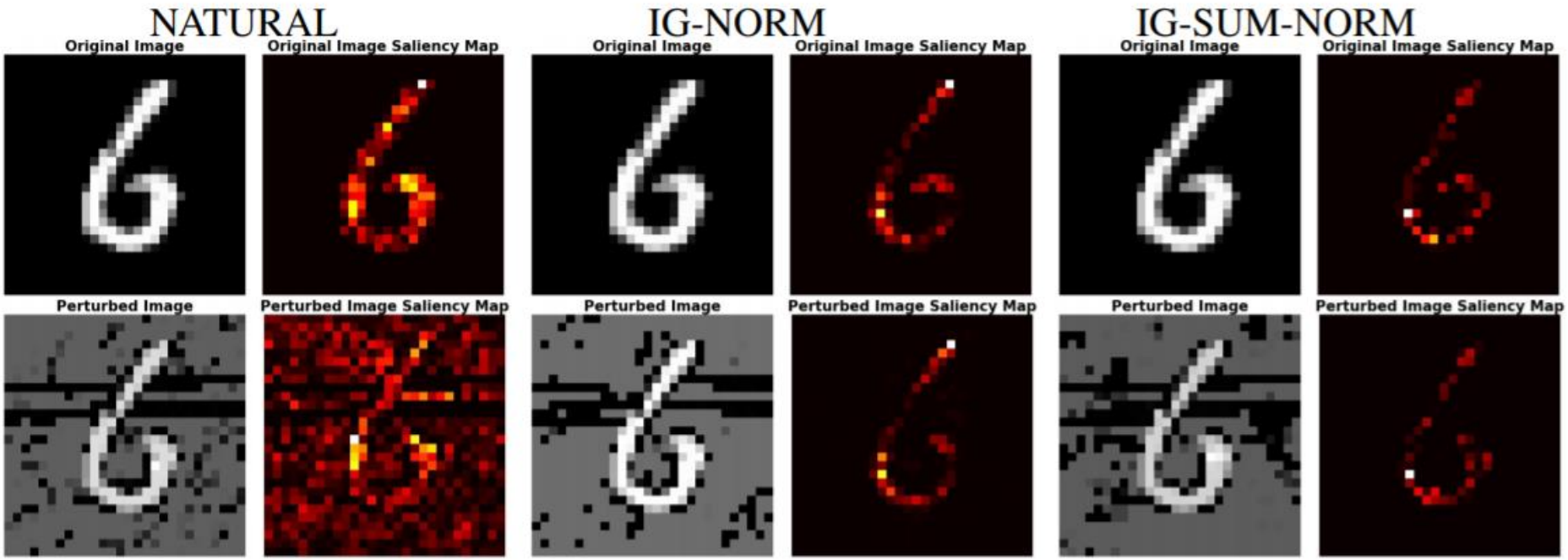
$$\text{IG-SUM-NORM} = \max_{\boldsymbol{x}' \in \Delta(\boldsymbol{x})} \left|\left|\mathrm{IG}(\boldsymbol{x}, \boldsymbol{x}')\right|\right|_1 + \mathrm{sum}(\mathrm{IG}(\boldsymbol{x}, \boldsymbol{x}'))$$

Experiments: Qualitative



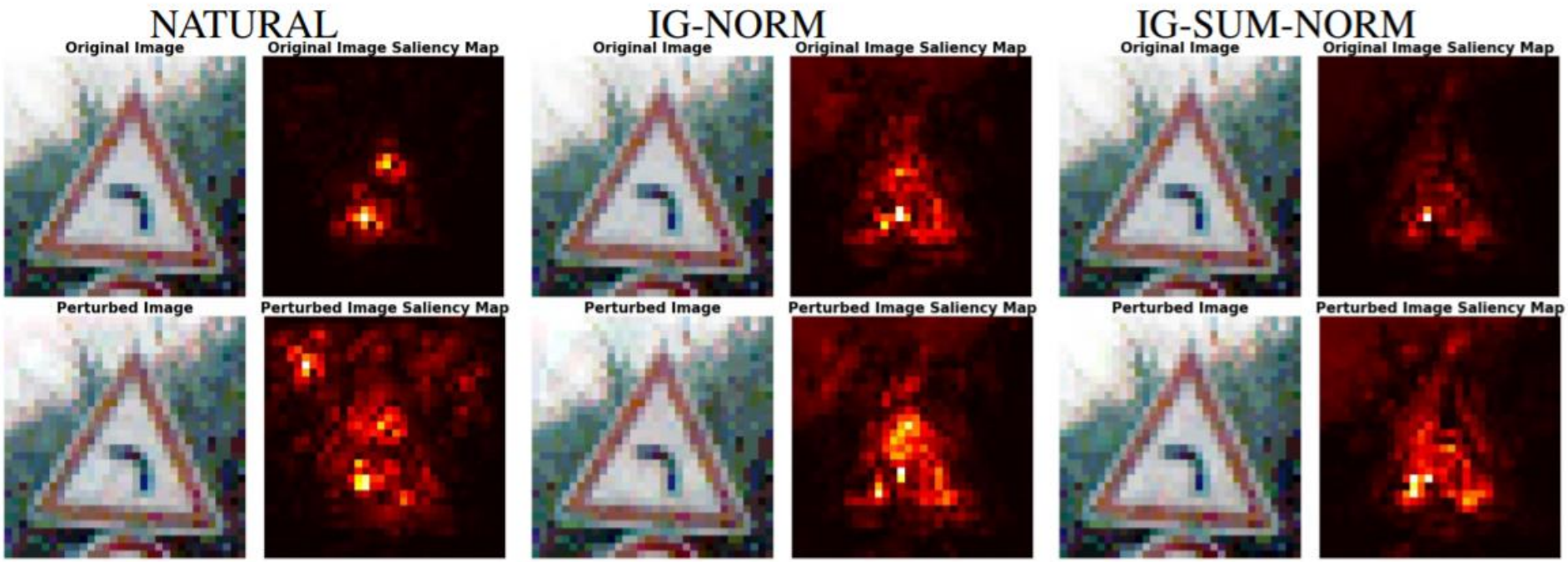
Flower dataset

Experiments: Qualitative



MNIST dataset

Experiments: Qualitative

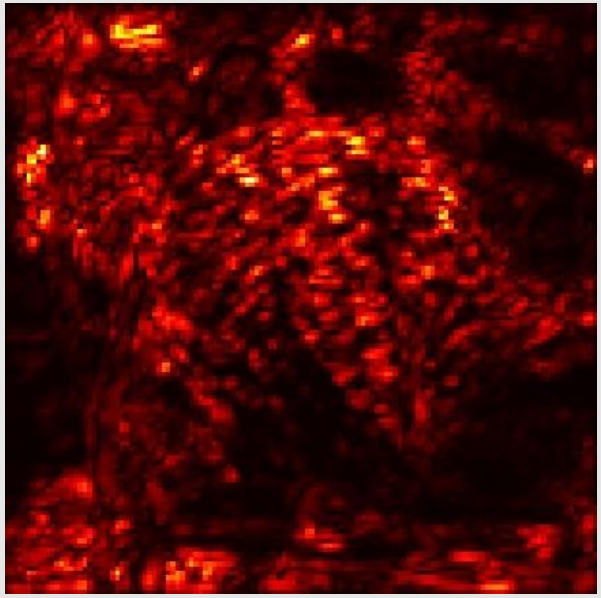


GTSRB dataset

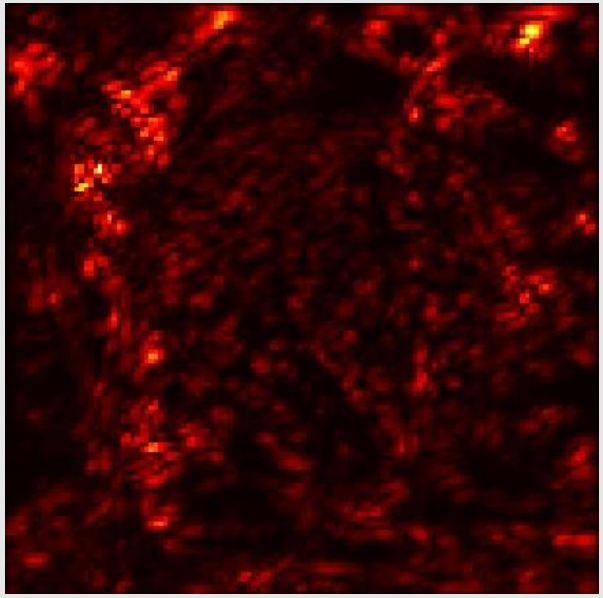
Experiments: Quantitative

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

Original Image Attribution Map

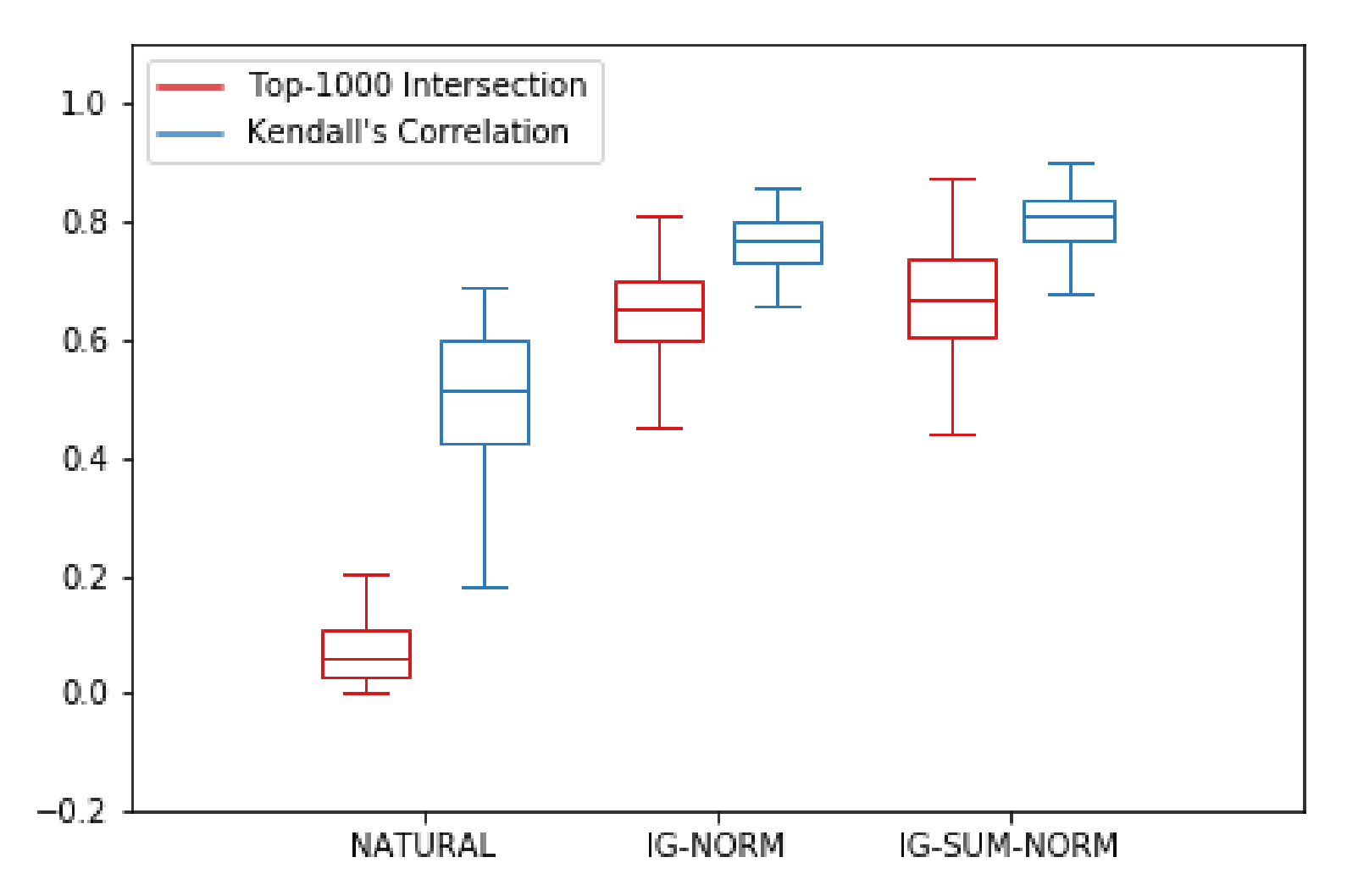


Perturbed Image Attribution Map

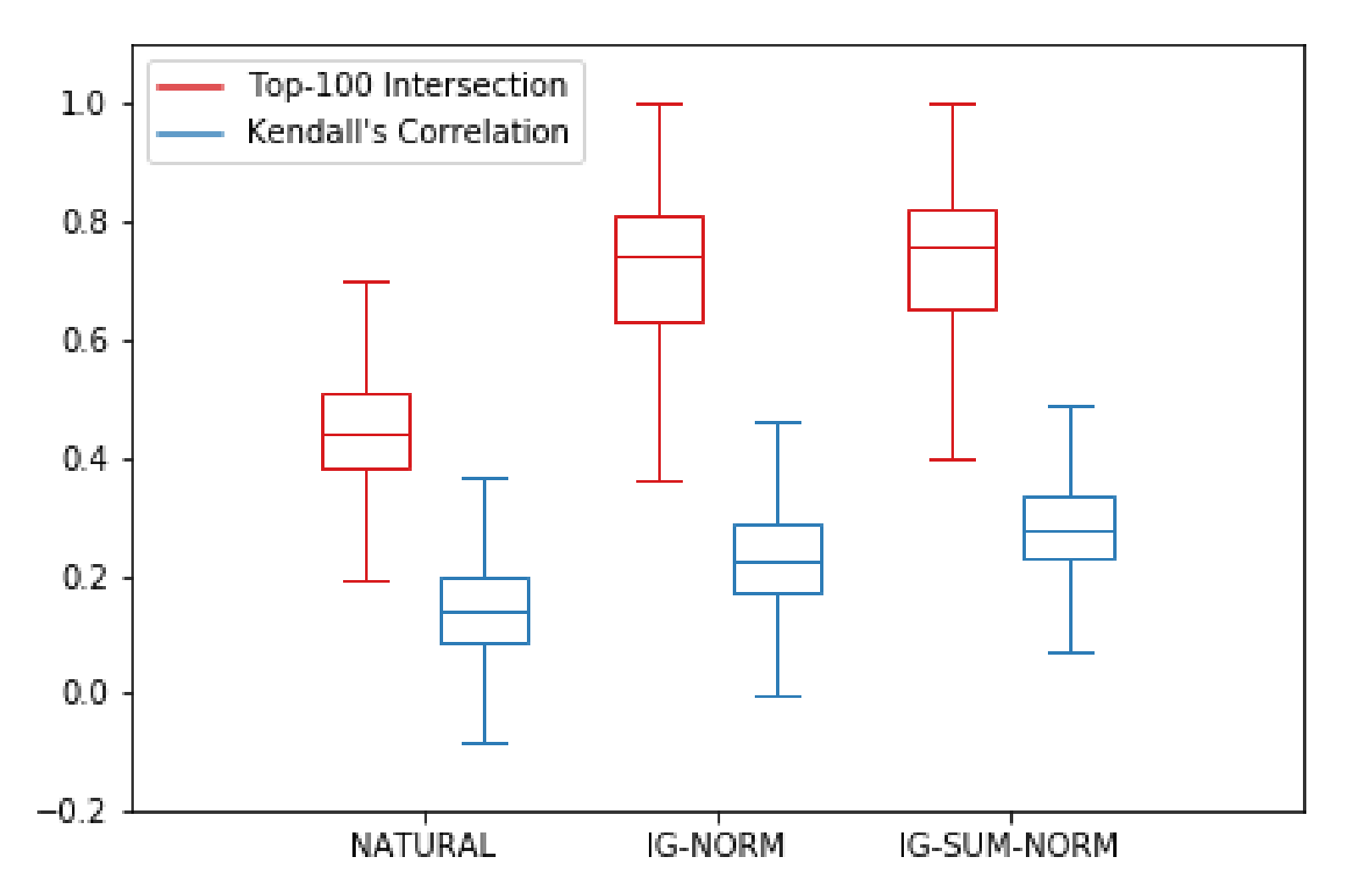


Top-1000 Intersection: 0.1%
Kendall's Correlation: 0.2607

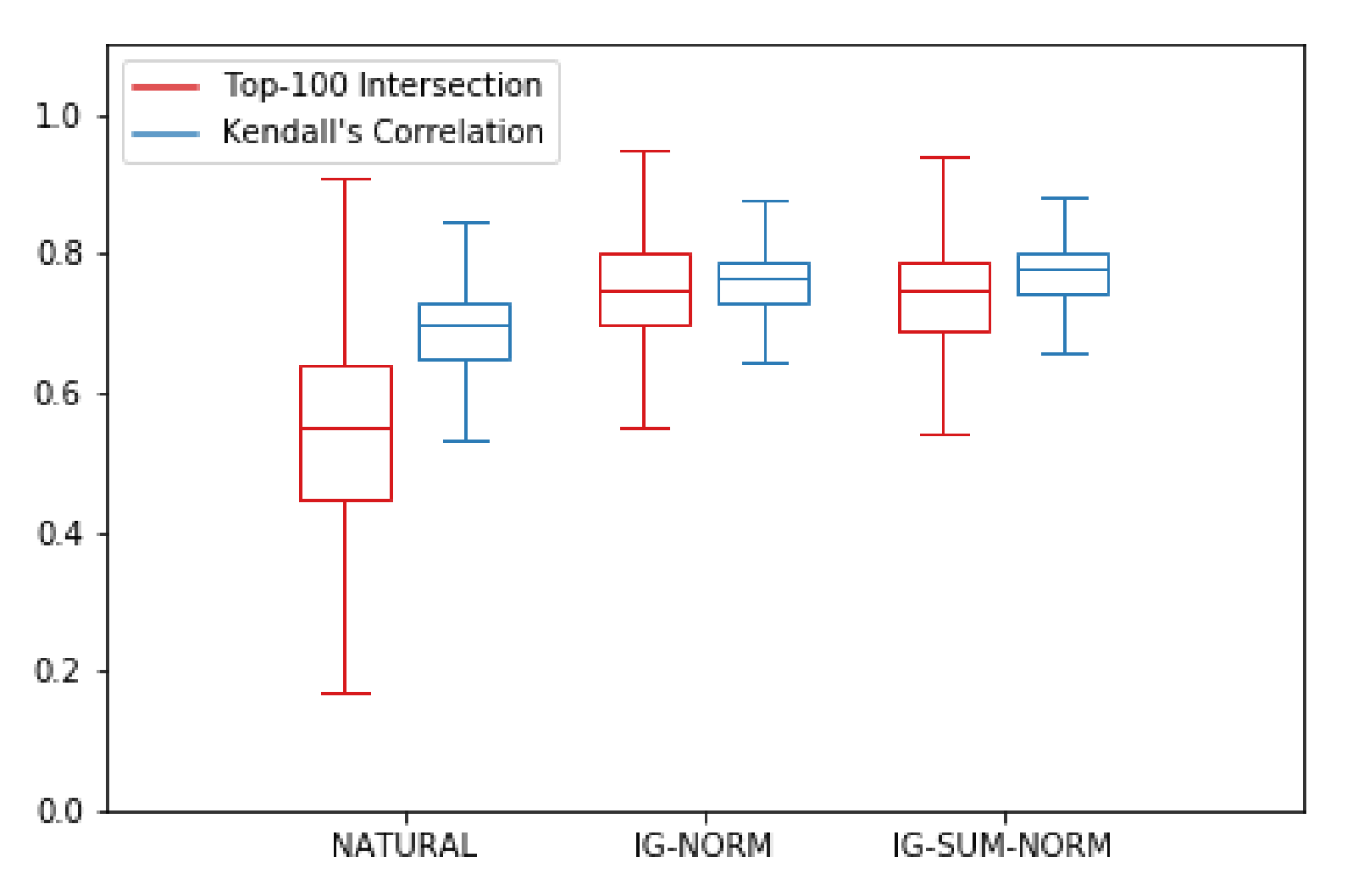
Result on Flower dataset



Result on MINST dataset



Result on GTSRB dataset



Connection to Robust Prediction

- RAR

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

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

- If $\lambda = 1$ and $s(\cdot) = sum(\cdot)$, then RAR becomes the Adversarial Training objective for robust prediction

$$\min_\theta \mathbb{E}\left[\max_{\boldsymbol{x}' \in N(\boldsymbol{x}, \epsilon)} l(\boldsymbol{x}', y; \theta)\right]$$

simply by the Completeness of IG

Towards Deep Learning Models Resistant to Adversarial Attacks.
Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, Adrian Vladu. ICML 2017.

When the two coincide?

- Theorem: For the special case of one-layer neural networks (linear function), the robust attribution instantiation ($s(\cdot) = \|\cdot\|_1$) and the robust prediction instantiation ($s(\cdot) = \mathrm{sum}(\cdot)$) coincide, and both reduce to soft max-margin training.

Connection to Robust Prediction

- RAR

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

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

- If $\lambda = \lambda'/\epsilon^q$ and $s(\cdot) = \|\cdot\|_1^q$ with approximate IG, then RAR becomes the Input Gradient Regularization for robust prediction

$$\min_\theta \mathbb{E}\left[l(\boldsymbol{x}, y; \theta) + \lambda' \|\nabla_{\boldsymbol{x}}\, l(\boldsymbol{x}, y; \theta)\|_q^q\right]$$

Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients.
Andrew Slavin Ross and Finale Doshi-Velez. AAAI 2018.


THANK YOU!