Robust Attribution Regularization

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

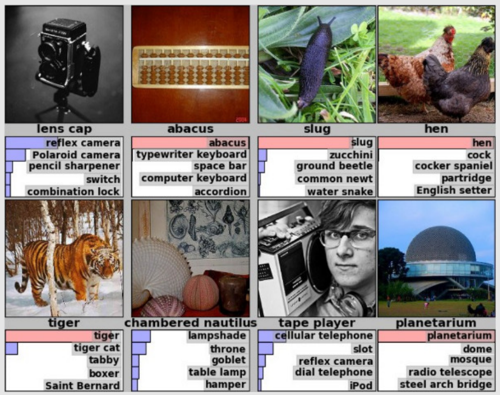
[1]University of Wisconsin-Madison [2]Google [3]XaiPient

NeurIPS'2019

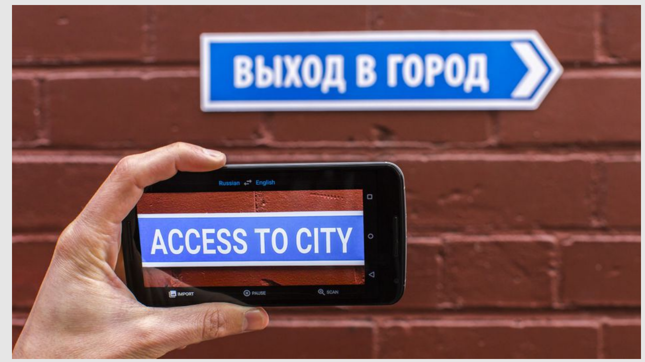
*Equal contribution
† Work done while at UW-Madison

Machine Learning Progress

- Significant progress in Machine Learning



Computer vision



Machine translation

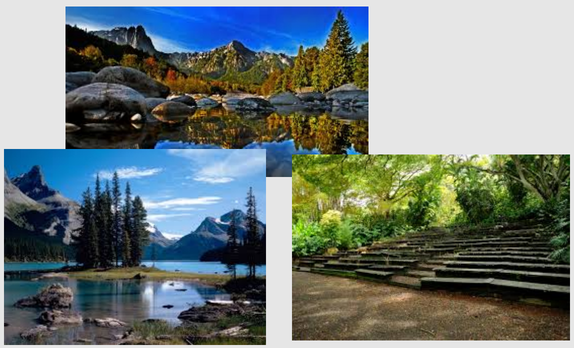
Game Playing

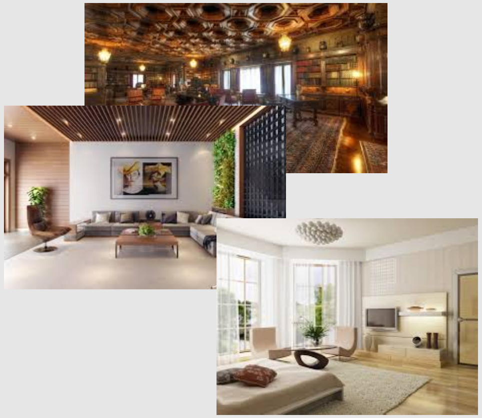
Medical Imaging

Key Engine Behind the Success

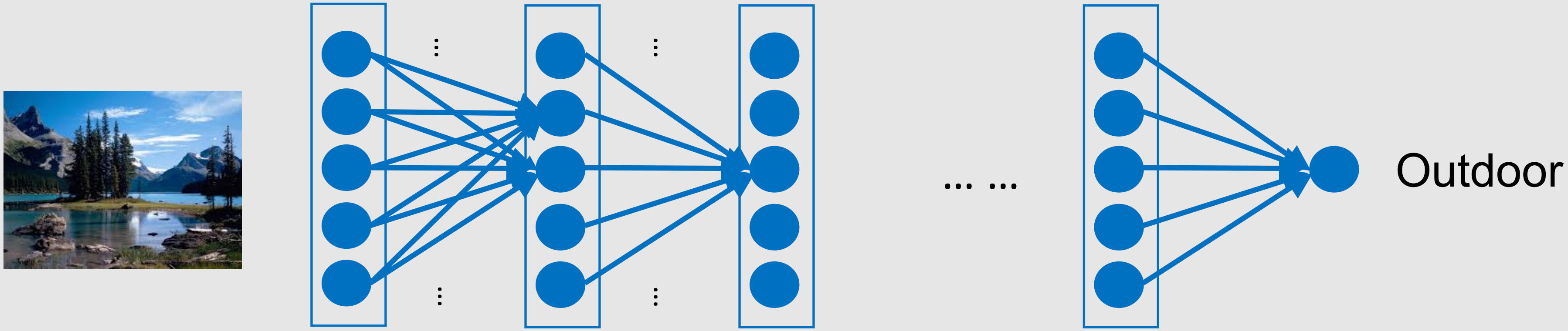
- Training Deep Neural Networks: $y = f(x; W)$
  - Given training data $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$
  - Try to find W such that the network fits the data

Outdoor



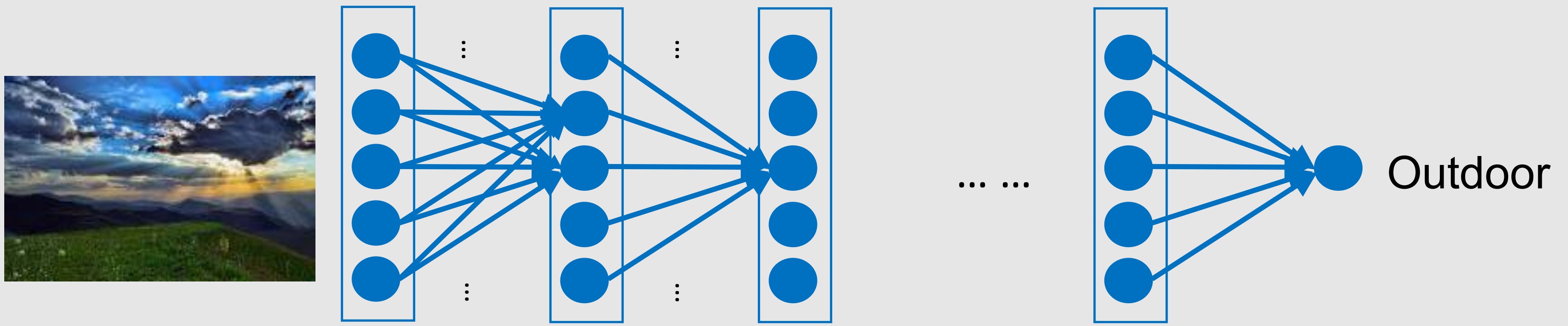


Indoor



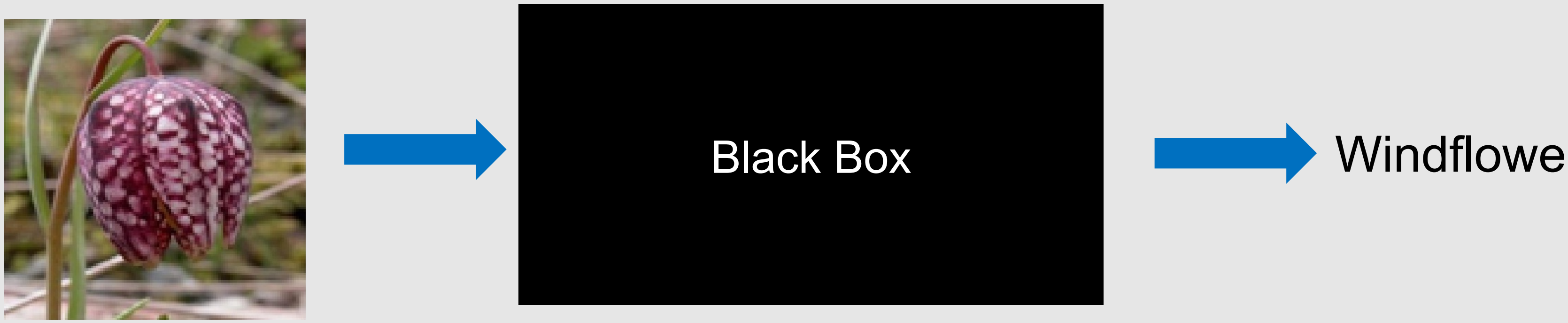
Key Engine Behind the Success

- Using Deep Neural Networks: $y = f(x; W)$
  - Given a new test point x
  - Predict $y = f(x; W)$

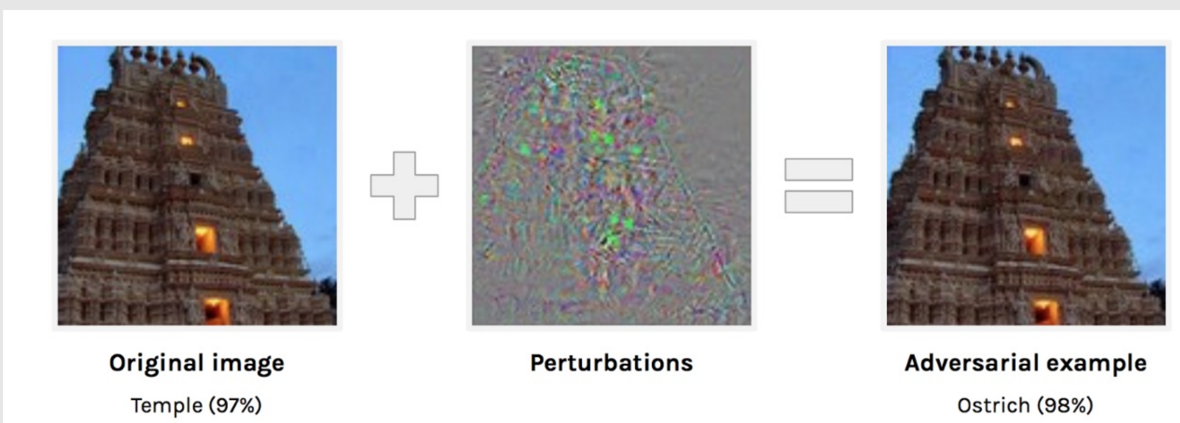


Challenges

- Blackbox: not too much understanding/interpretation

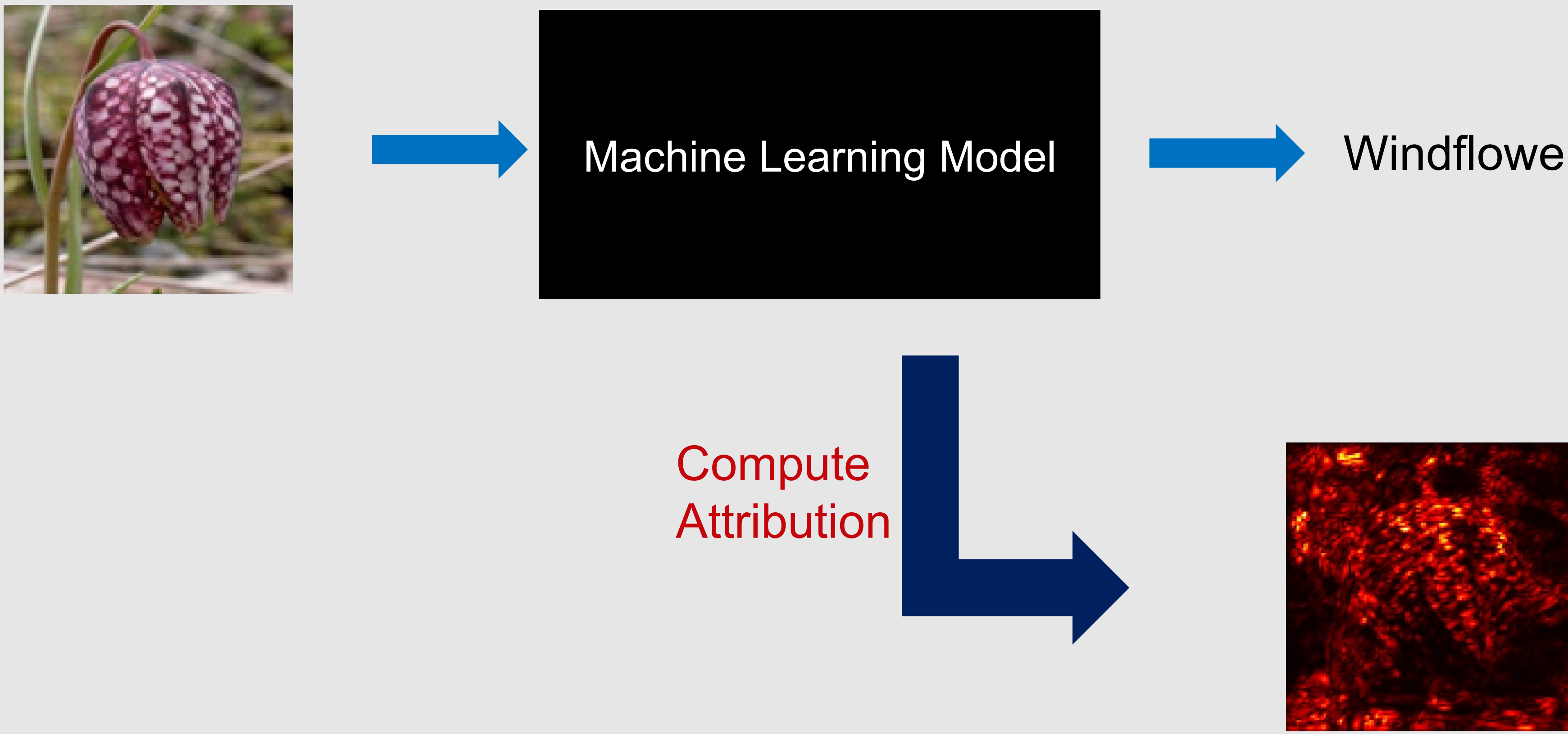


- Vulnerable to adversaries



Interpretable Machine Learning

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



Integrated Gradient: Axiomatic Approach

Overview

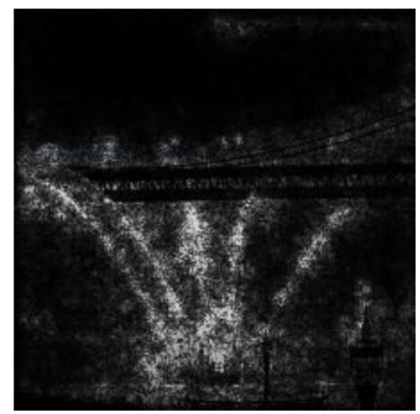
- List desirable criteria (axioms) for an attribution method
- Establish a uniqueness result: only this method satisfies these desirable criteria

- 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: Definition

$$IG(input,base) = (input - baseline) * \int_{0-1} \nabla F(\alpha * input + (1-\alpha) * baseline)\, d\alpha$$



Integrated Gradient: Example Results

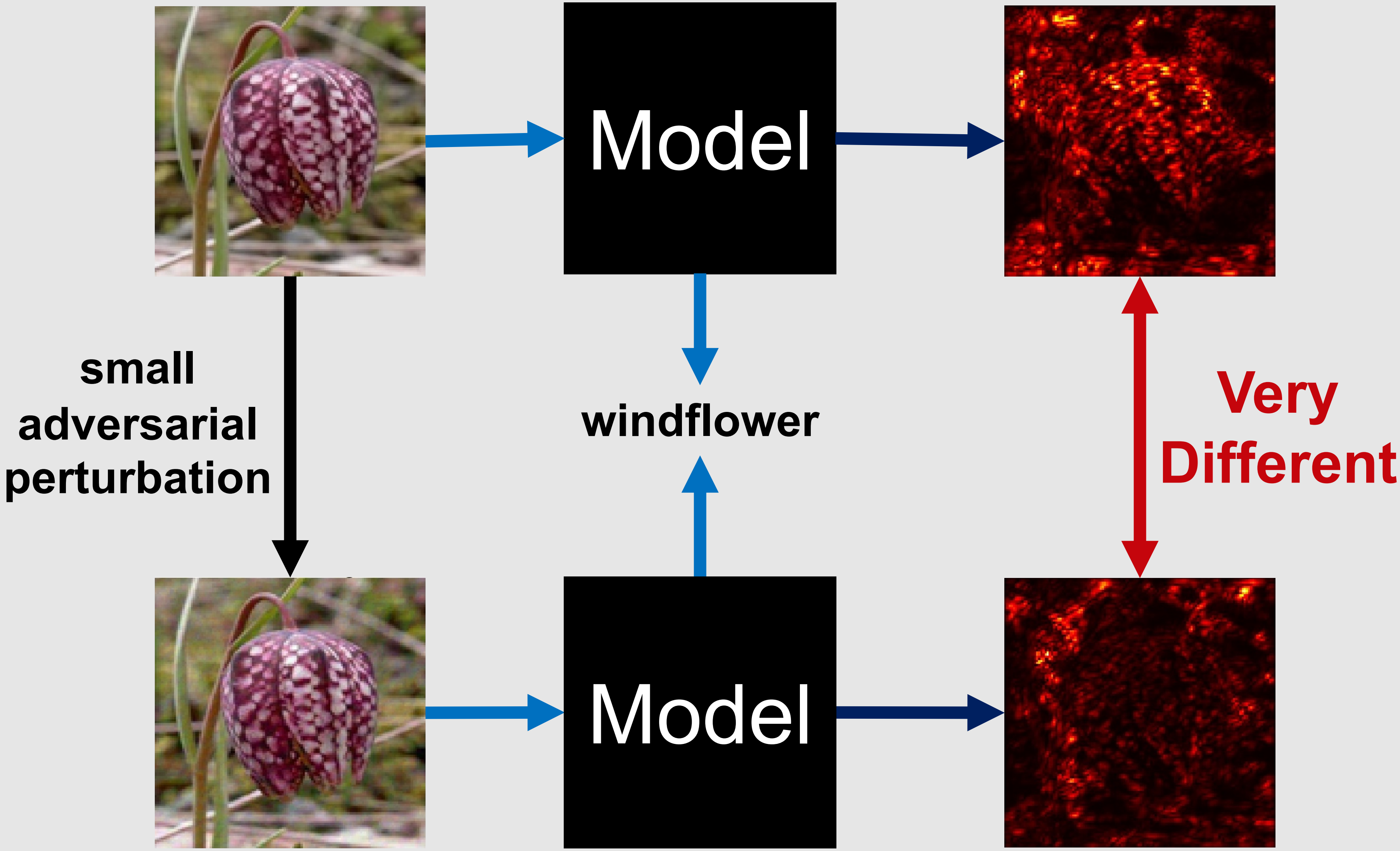
Integrated Gradient: Axioms

- **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

Attribution is Fragile

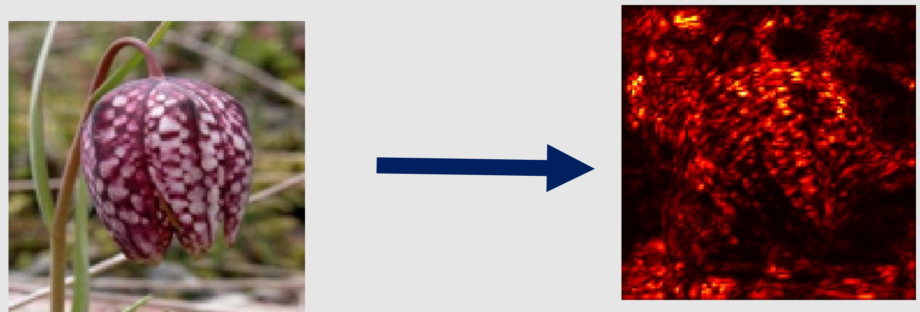


Interpretation of Neural Networks is Fragile.
Amirata Ghorbani, Abubakar Abid, James Zou. AAAI 2019.

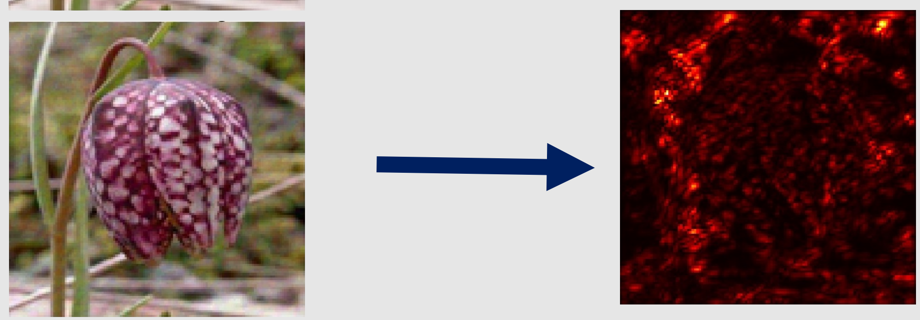
Robust Prediction Correlates with Robust Attribution: Why?

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

original image,
normally trained model



perturbed image,
normally trained model

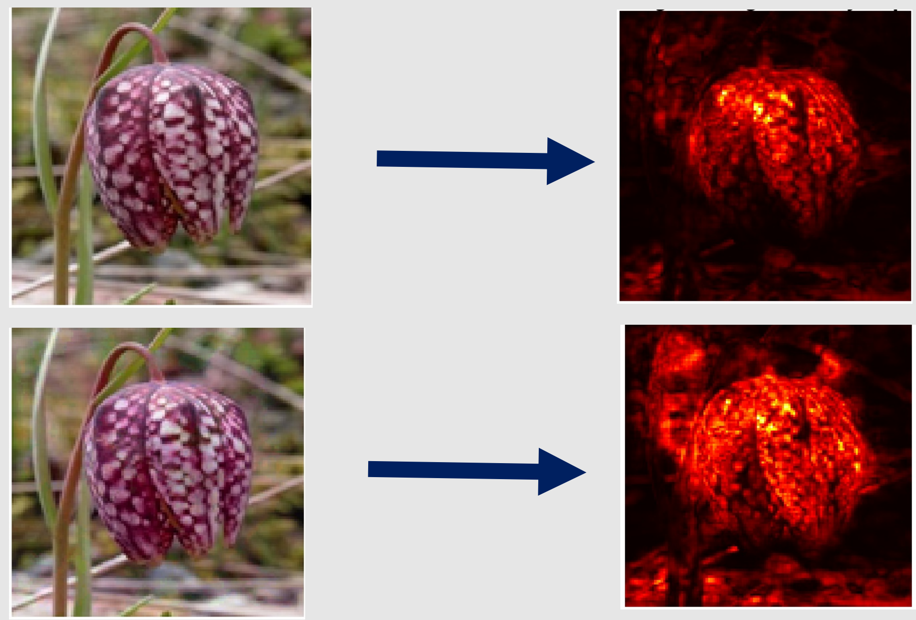


Robust Prediction Correlates with Robust Attribution: Why?

- 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

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

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

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

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

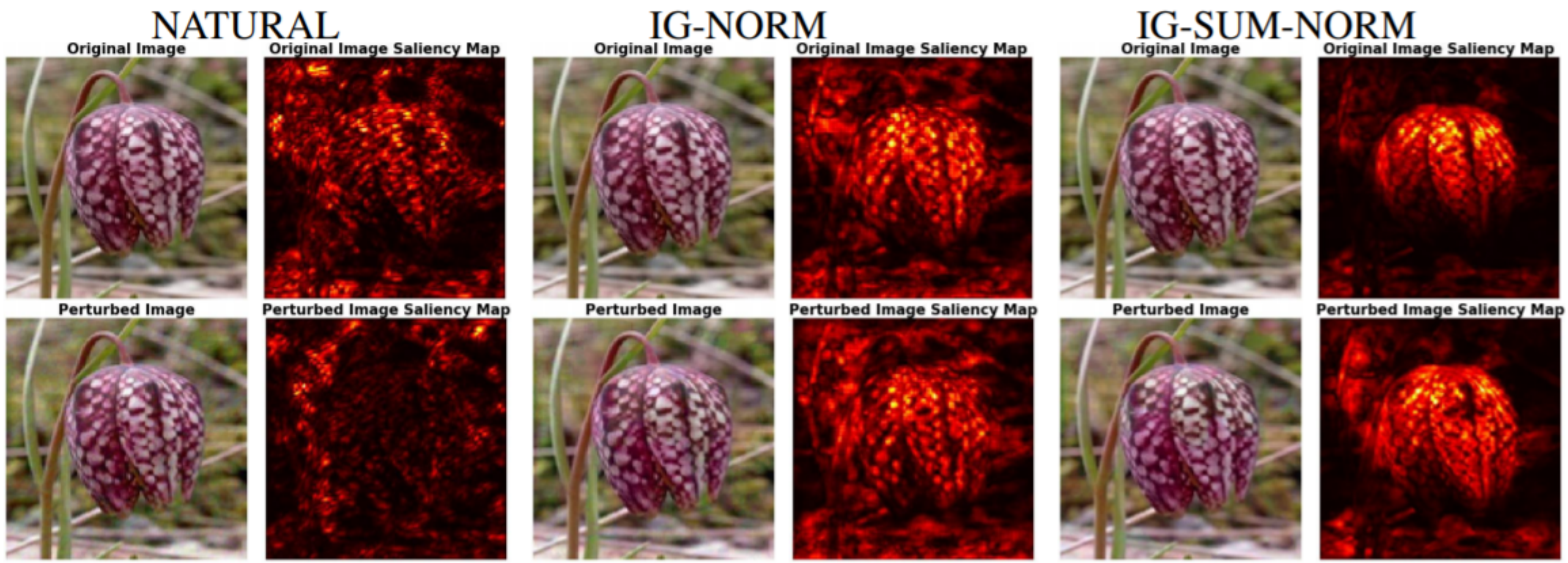
$$\text{RAR} = \max_{\boldsymbol{x}' \in \Delta(\boldsymbol{x})} s(\text{IG}(\boldsymbol{x}, \boldsymbol{x}'))$$

- Two instantiations:

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

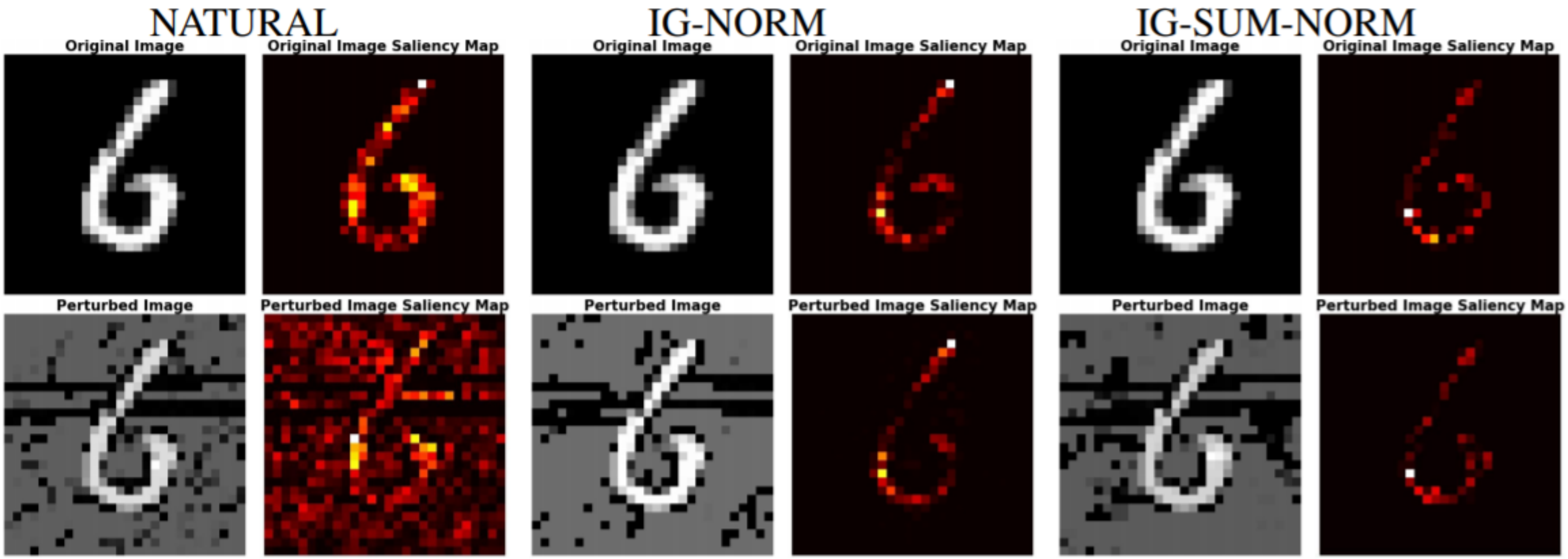
$$\text{IG-SUM-NORM} = \max_{\boldsymbol{x}' \in \Delta(\boldsymbol{x})} \left|\left|\text{IG}(\boldsymbol{x}, \boldsymbol{x}')\right|\right|_1 + \text{sum}(\text{IG}(\boldsymbol{x}, \boldsymbol{x}'))$$

Experiments: Qualitative



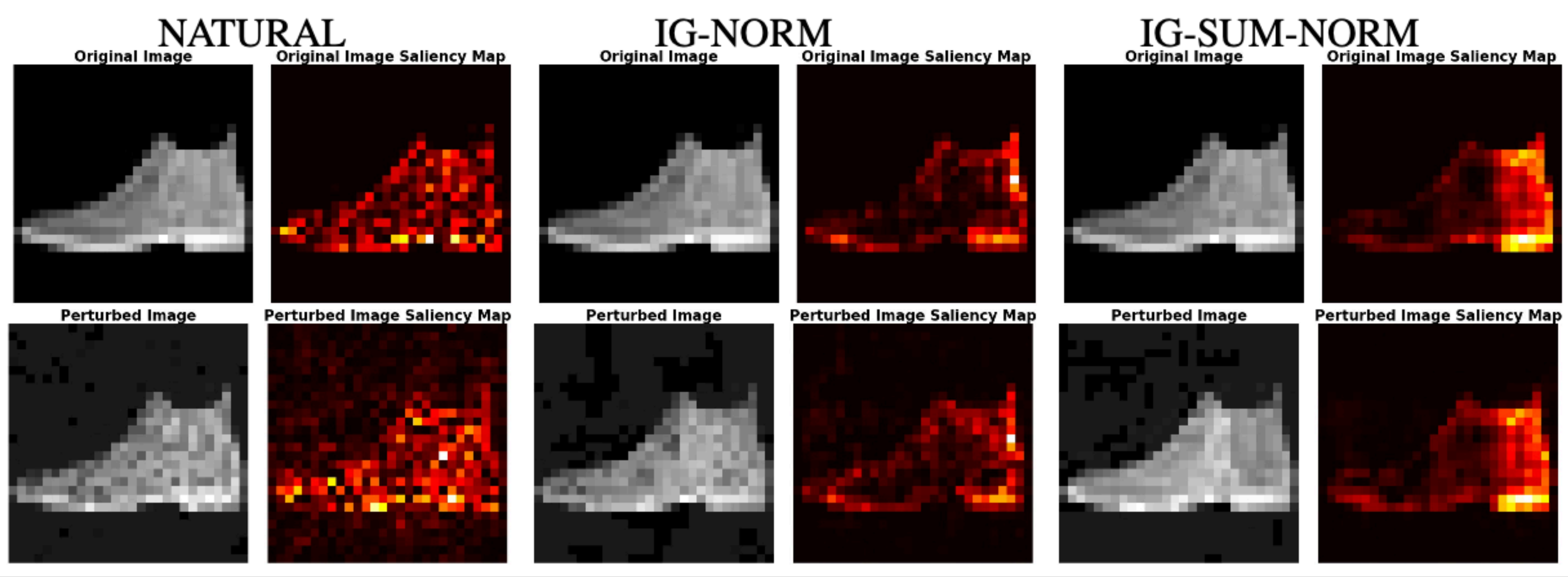
Flower dataset

Experiments: Qualitative



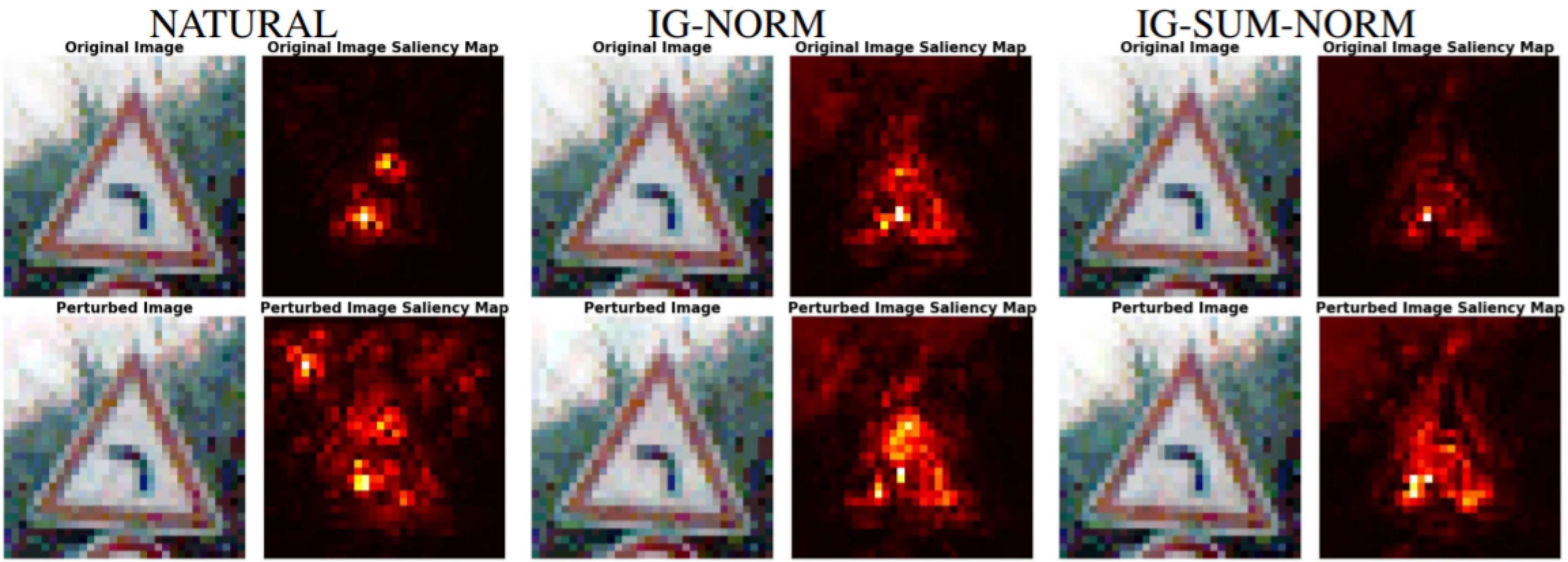
MNIST dataset

Experiments: Qualitative



Fashion-MNIST dataset

Experiments: Qualitative

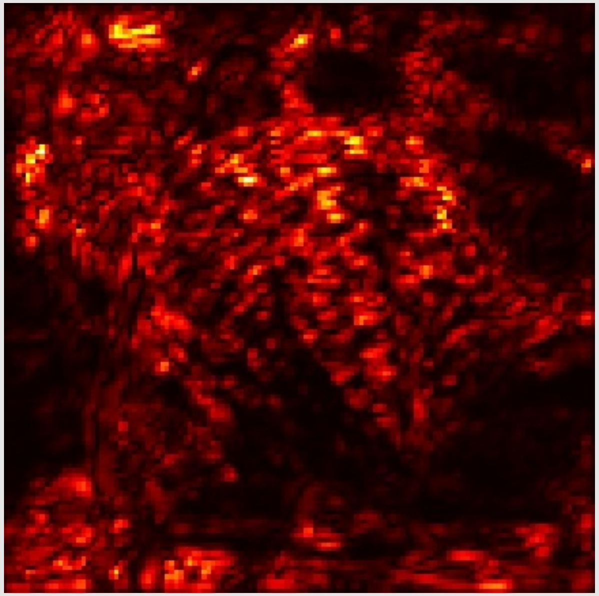


GTSRB dataset

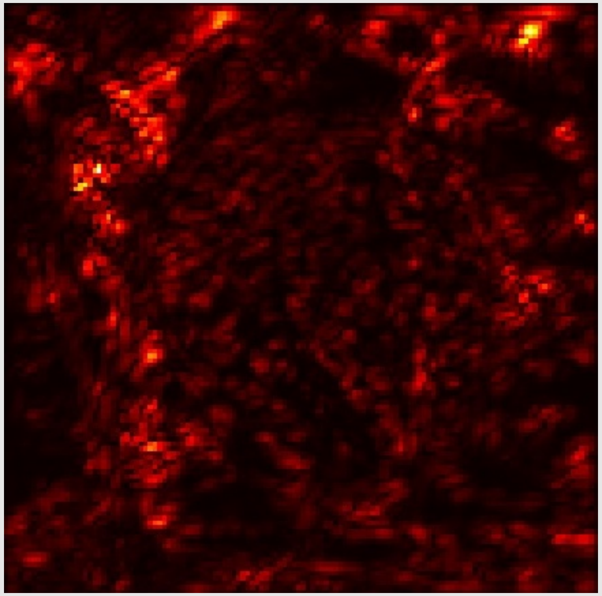
Experiments: Quantitative

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

Original Image Attribution Map

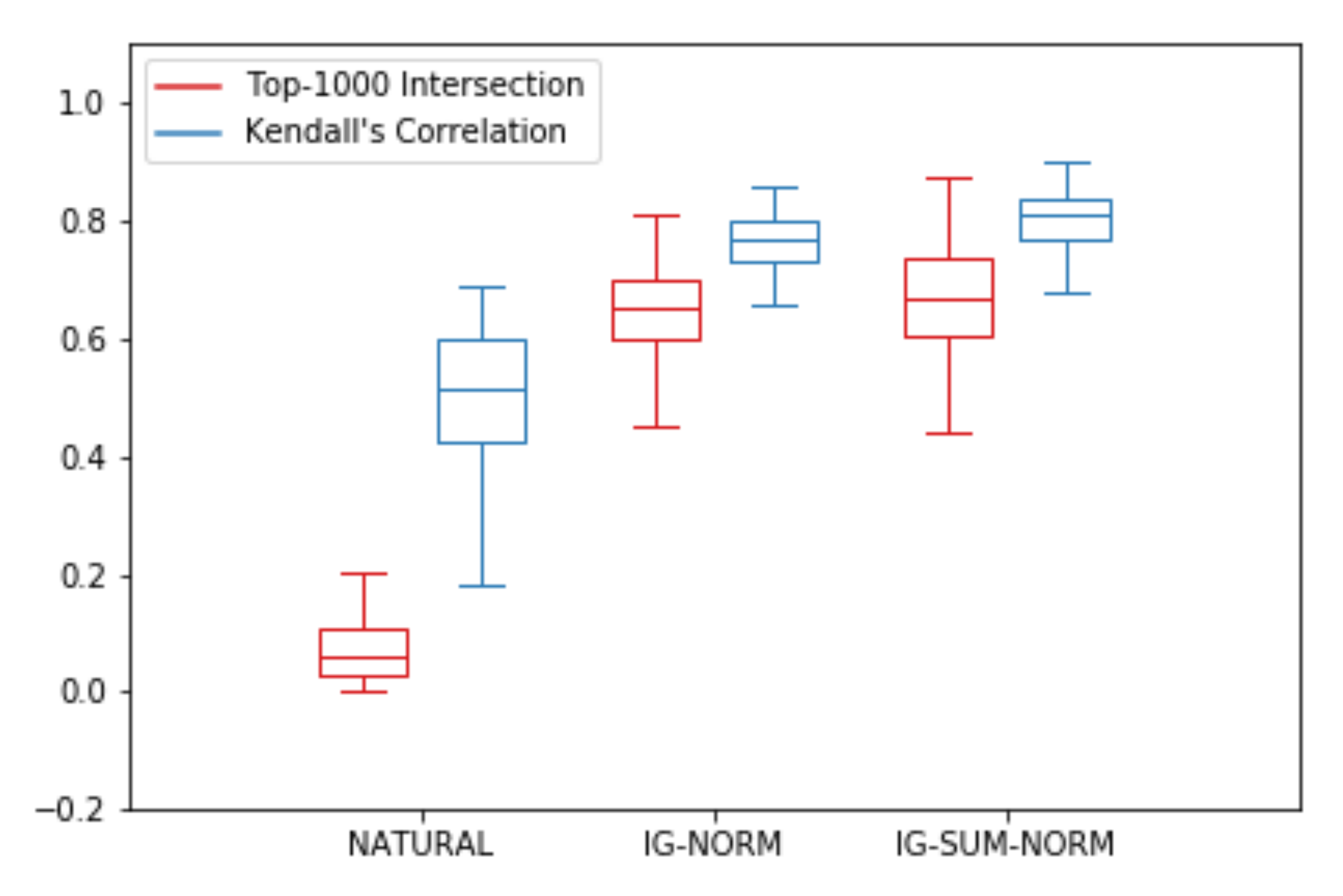


Perturbed Image Attribution Map

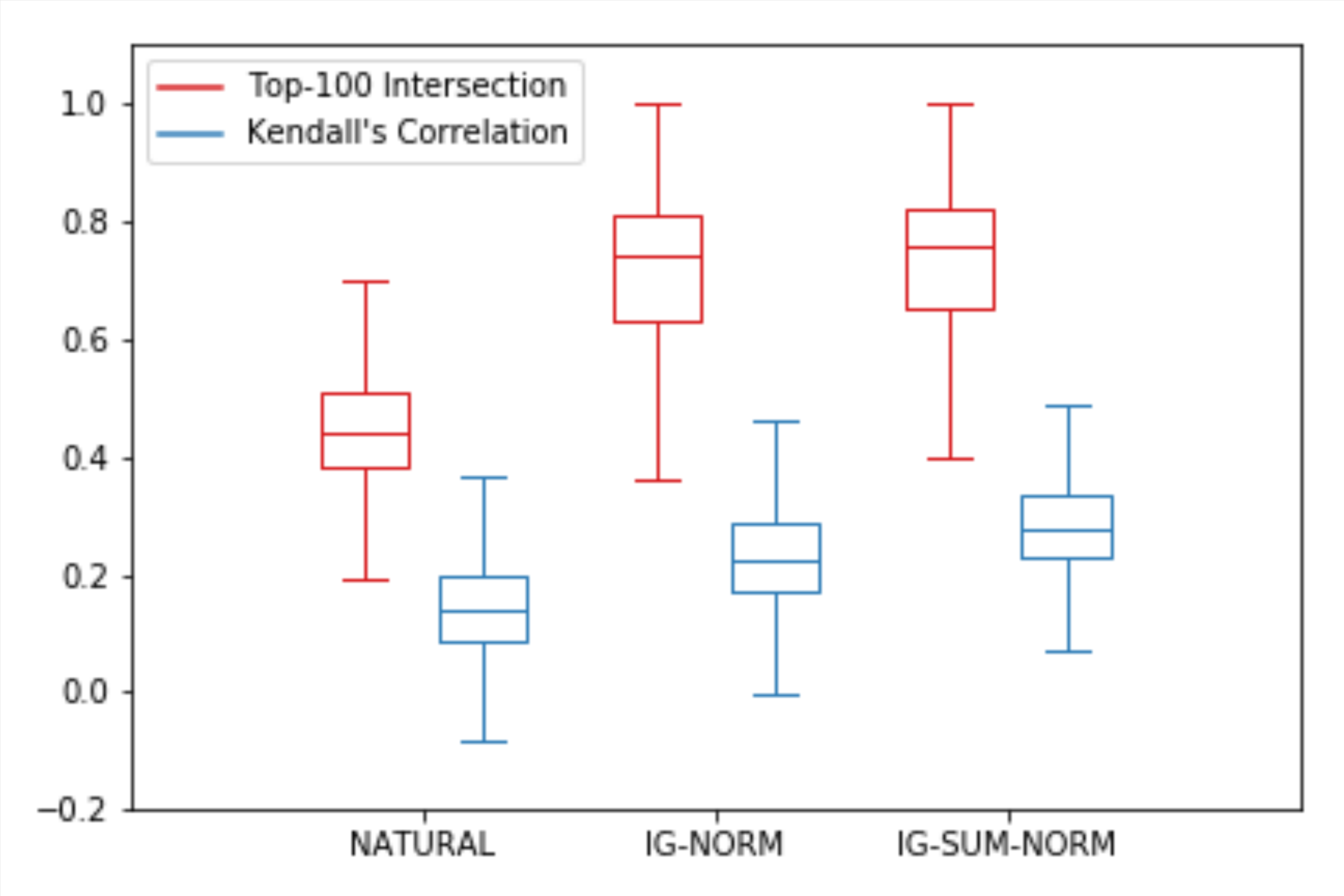


Top-1000 Intersection: 0.1%
Kendall's Correlation: 0.2607

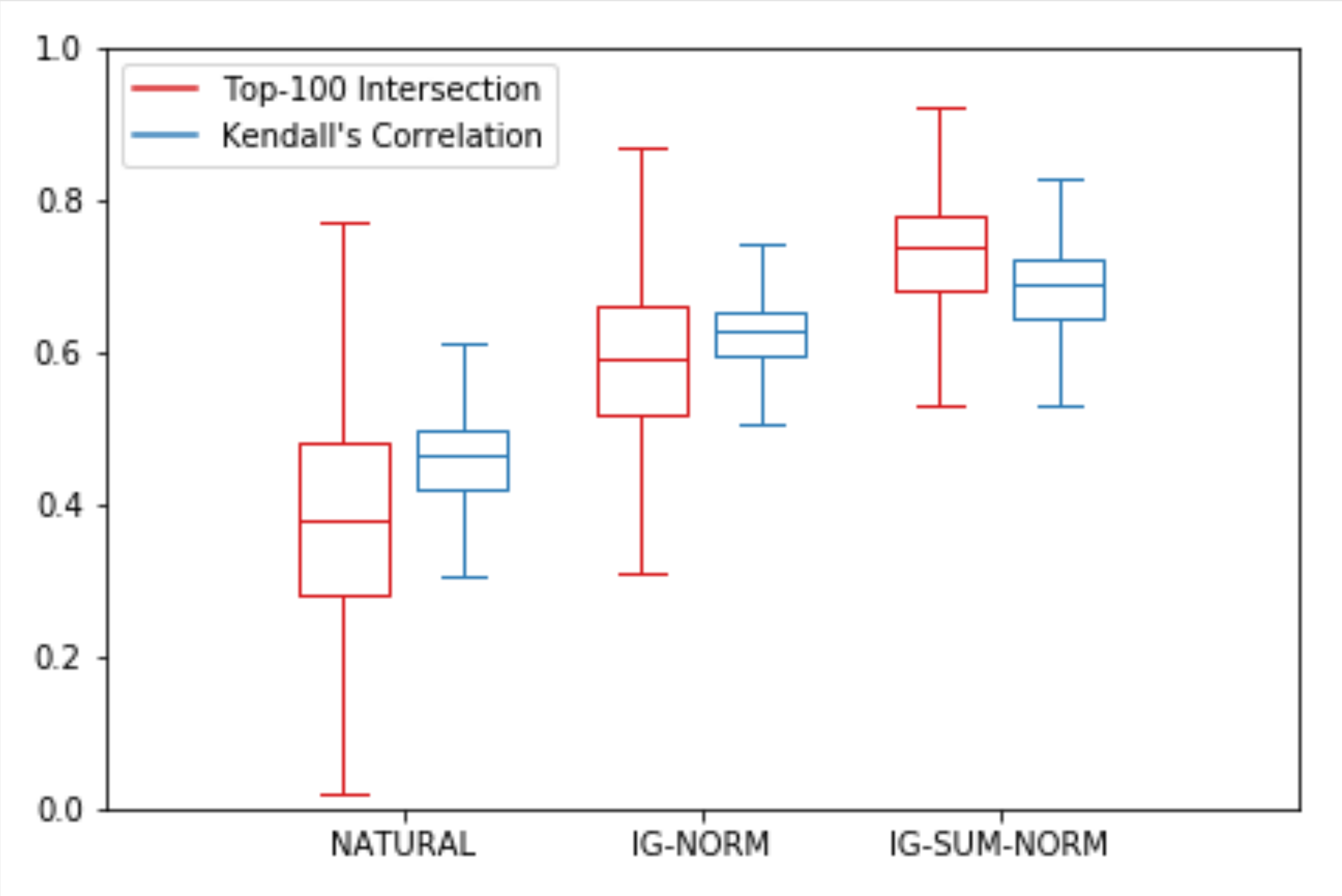
Result on Flower dataset



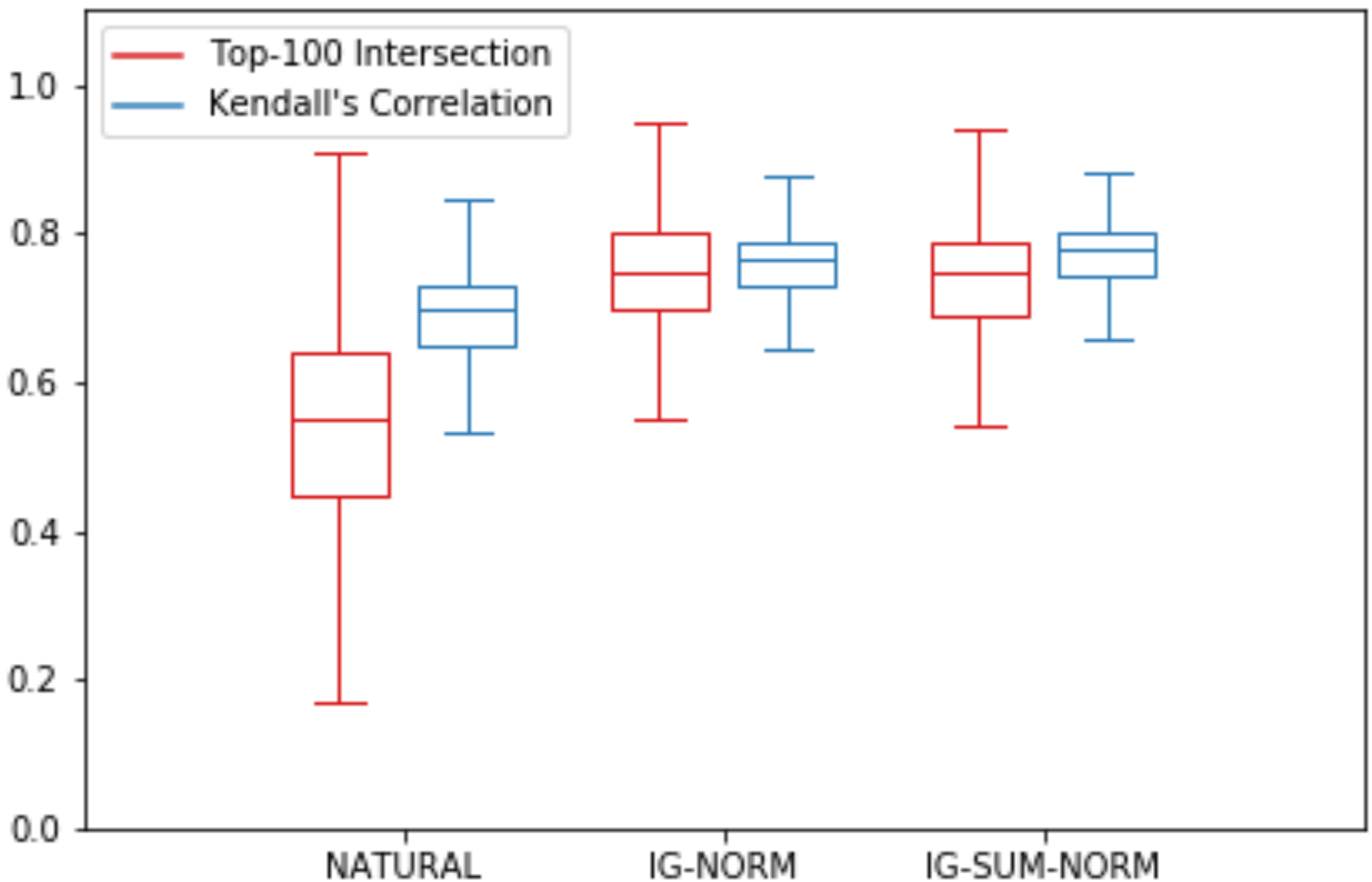
Result on MINST dataset



Result on Fashion-MINST dataset



Result on GTSRB dataset



Prediction Accuracy of Different Models

Dataset	Approach	Accuracy
MNIST	NATURAL	99.17%
	IG-NORM	98.74%
	IG-SUM-NORM	98.34%
Fashion-MNIST	NATURAL	90.86%
	IG-NORM	85.13%
	IG-SUM-NORM	85.44%
GTSRB	NATURAL	98.57%
	IG-NORM	97.02%
	IG-SUM-NORM	95.68%
Flower	NATURAL	86.76%
	IG-NORM	85.29%
	IG-SUM-NORM	82.35%

Connection to Robust Prediction

- RAR

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

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

- 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

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

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

- 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.

Discussion

- Robust attribution leads to more human-aligned attribution.
- Robust attribution may help tackle spurious correlations.

THANK YOU!