Concise Explanations of Neural Networks using Adversarial Training

Prasad Chalasani (XaiPient)

ICML 2020

Co-authors:

Jiefeng Chen (UW Madison)

Amrita Roy Chowdhury (UW Madison)

Xi Wu (Google)

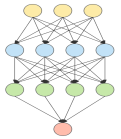
Somesh Jha (UW Madison, XaiPient)

Quick overview

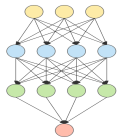
Deep Learning models: two key concerns

(1) Lack of explainability

Patient history

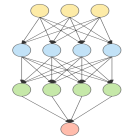


High risk of heart failure



P(COVID-19) = 0.85

Loan Application

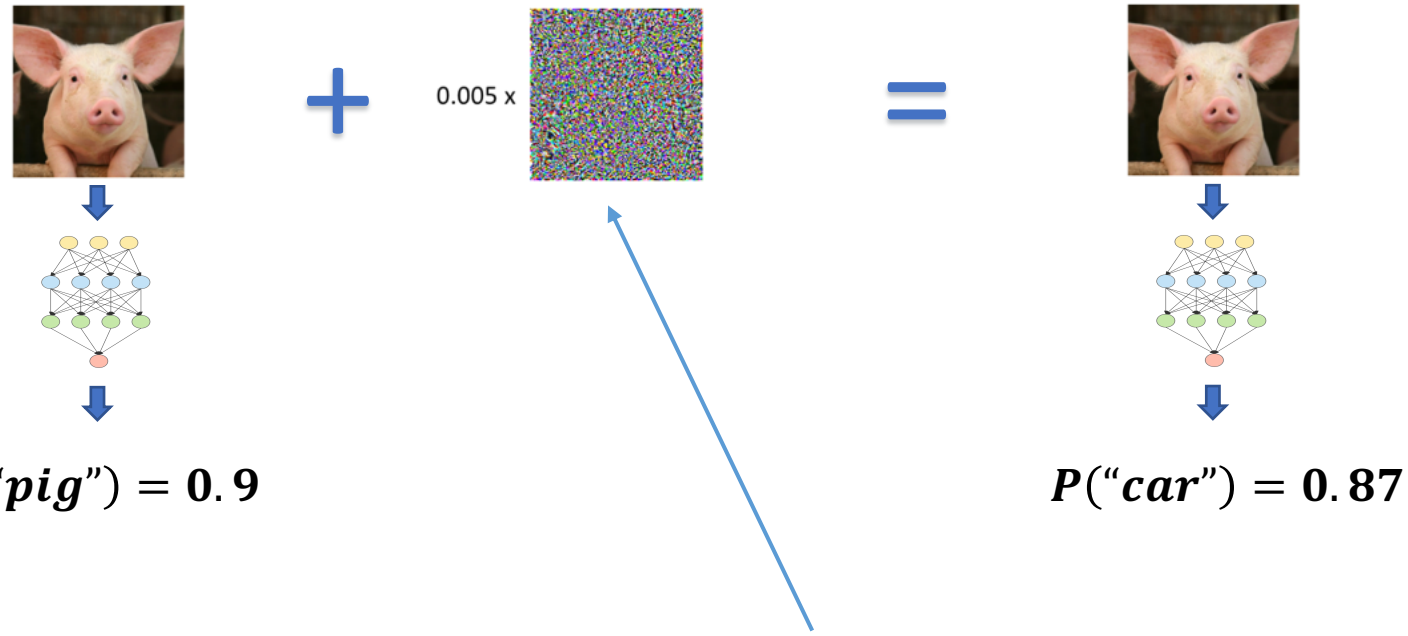


Score = 0.3

Why?

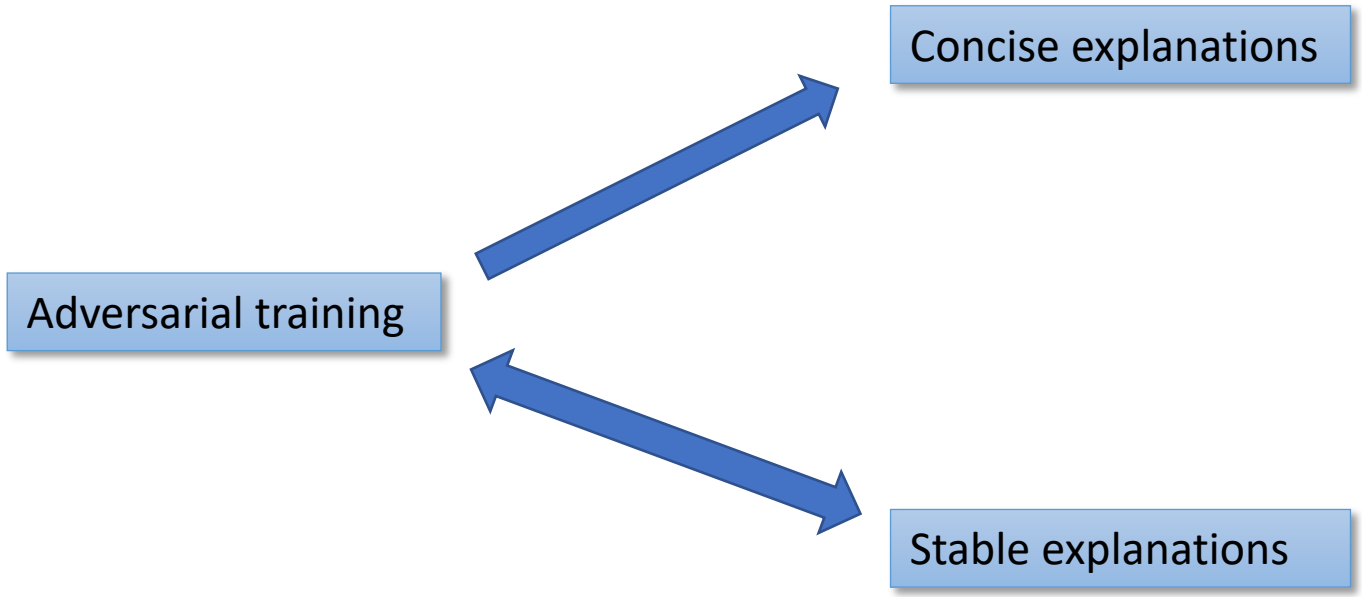
Deep Learning models: two key concerns

(2) Vulnerability to adversarial attacks



(Images from Madry et al, Gradient Science blog, http://gradientscience.org/intro_adversarial/)

Our results:
these two issues are related !



"Robust models have better explanations"

Deeper dive

Deep Learning models: two key concerns

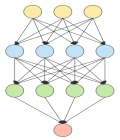
Explainability

(Trained) Neural Network: $\boldsymbol{x} \rightarrow F(\boldsymbol{x};\ \boldsymbol{w})$

Model / architecture

Weights / parameters

$\boldsymbol{x}$



$F(\boldsymbol{x};\ \boldsymbol{w})$

We'd like to "Explain" in terms of $\boldsymbol{x_i}$

Deep Learning models: two key concerns

Feature attribution methods

(IG, SHAP, LIME, DeepLIFT…)

Neural Network: $\boldsymbol{x} \rightarrow \ y = \ F(\boldsymbol{x})$ (omit $\boldsymbol{w}$ for brevity)

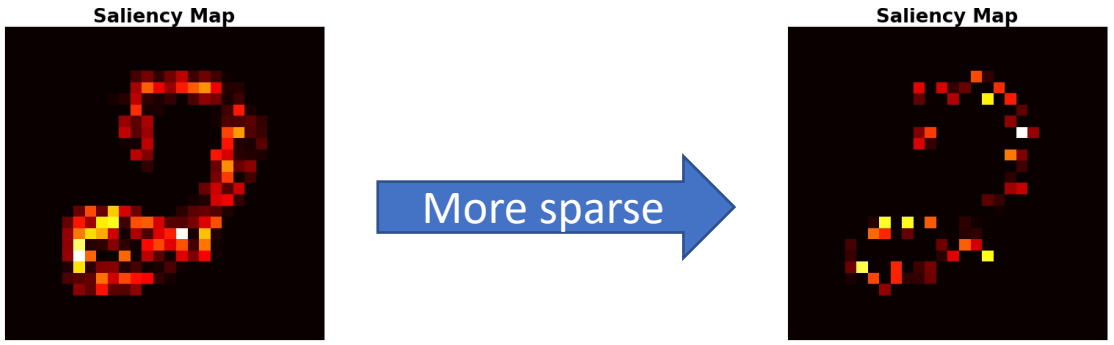
Attributions: $(\boldsymbol{x}, \boldsymbol{F}) \rightarrow \ A^F(\boldsymbol{x}\ )$

"explanation"

$A^F(\boldsymbol{x}\ )_i =$ "contribution" of $\boldsymbol{x}_i$ to $F(\boldsymbol{x})$

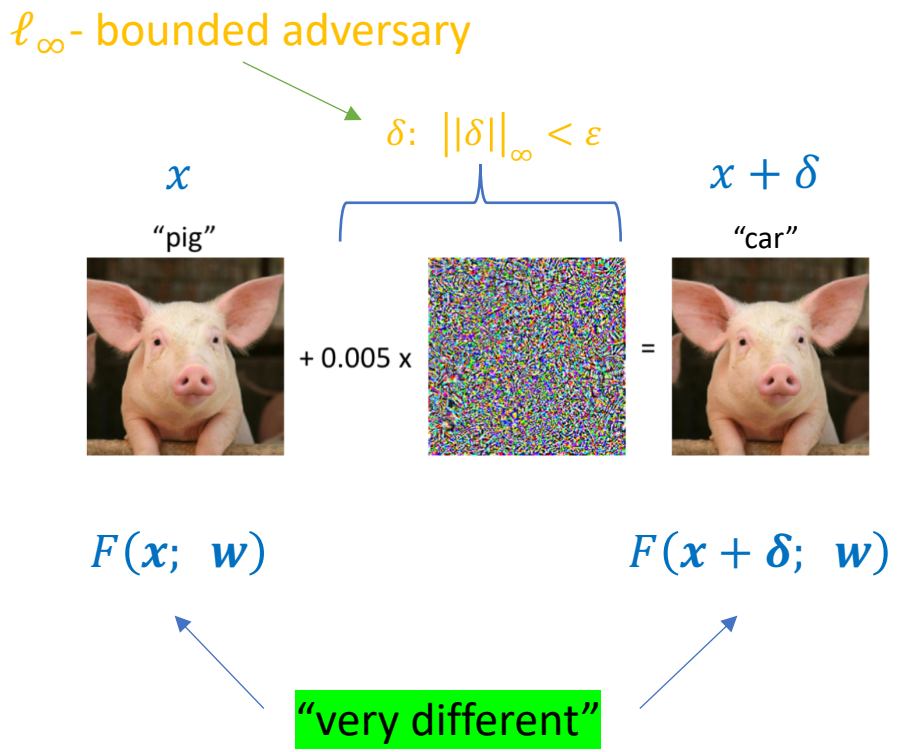
Desirable properties of **human-friendly** explanations

Conciseness	Stability
• Only focus on the truly relevant features: Sparse $A^F(\boldsymbol{x})$	• Should not change much when x changes slightly: Stable $A^F(\boldsymbol{x})$



Deep Learning models: two key concerns

Vulnerability to adversarial attacks



Deep Learning models: two key concerns

$\ell_\infty(\varepsilon)$-adversarial training

Instead of this:

Train $\boldsymbol{w}$ to minimize
Expected loss:

$$\min_{\boldsymbol{w}} \; E_{(\boldsymbol{x},y)\sim D} \; L(\boldsymbol{x}, y; \; \boldsymbol{w})$$

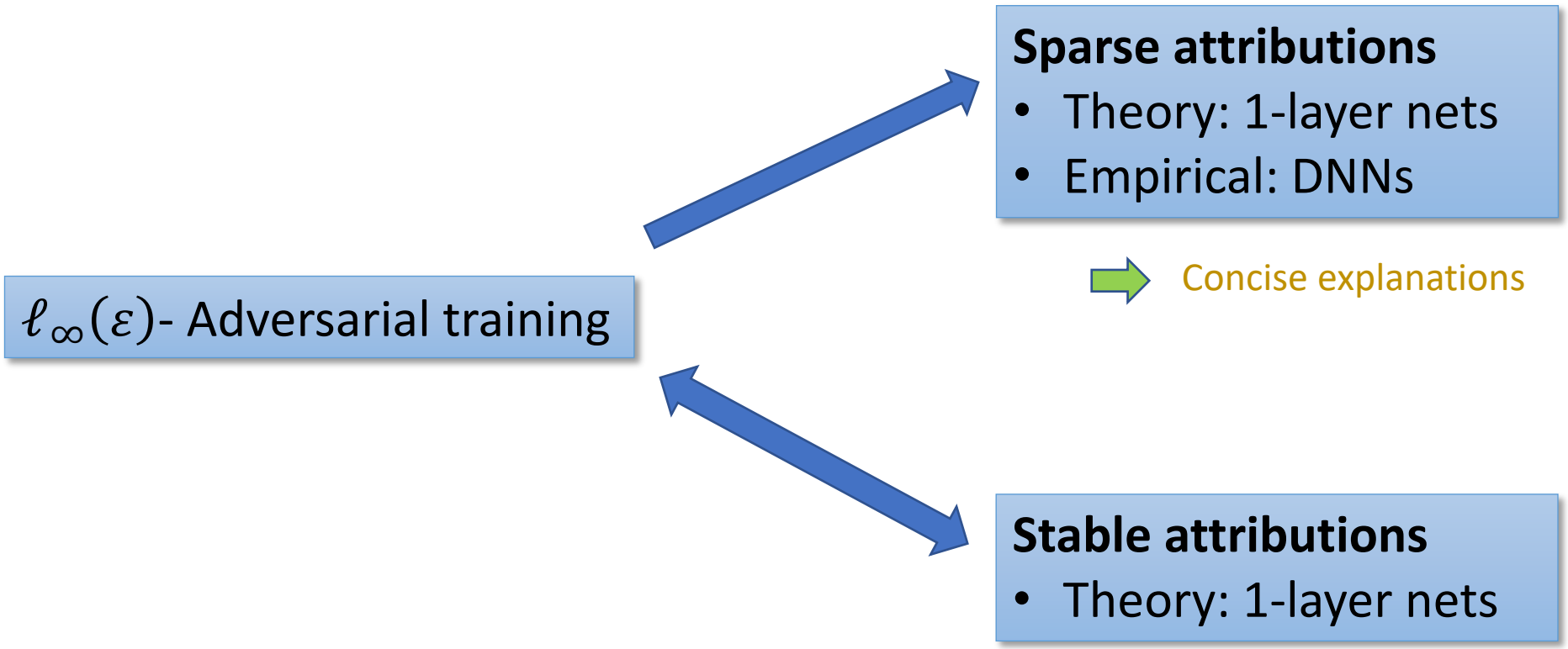
Do this:

Train $\boldsymbol{w}$ to minimize
Expected $\ell_\infty(\varepsilon)$ - Adversarial Loss:

$$\min_{\boldsymbol{w}} \; E_{(\boldsymbol{x},y)\sim D} \; \max_{||\delta||_\infty<\varepsilon} \; L(\boldsymbol{x}+\delta, y; \; \boldsymbol{w})$$

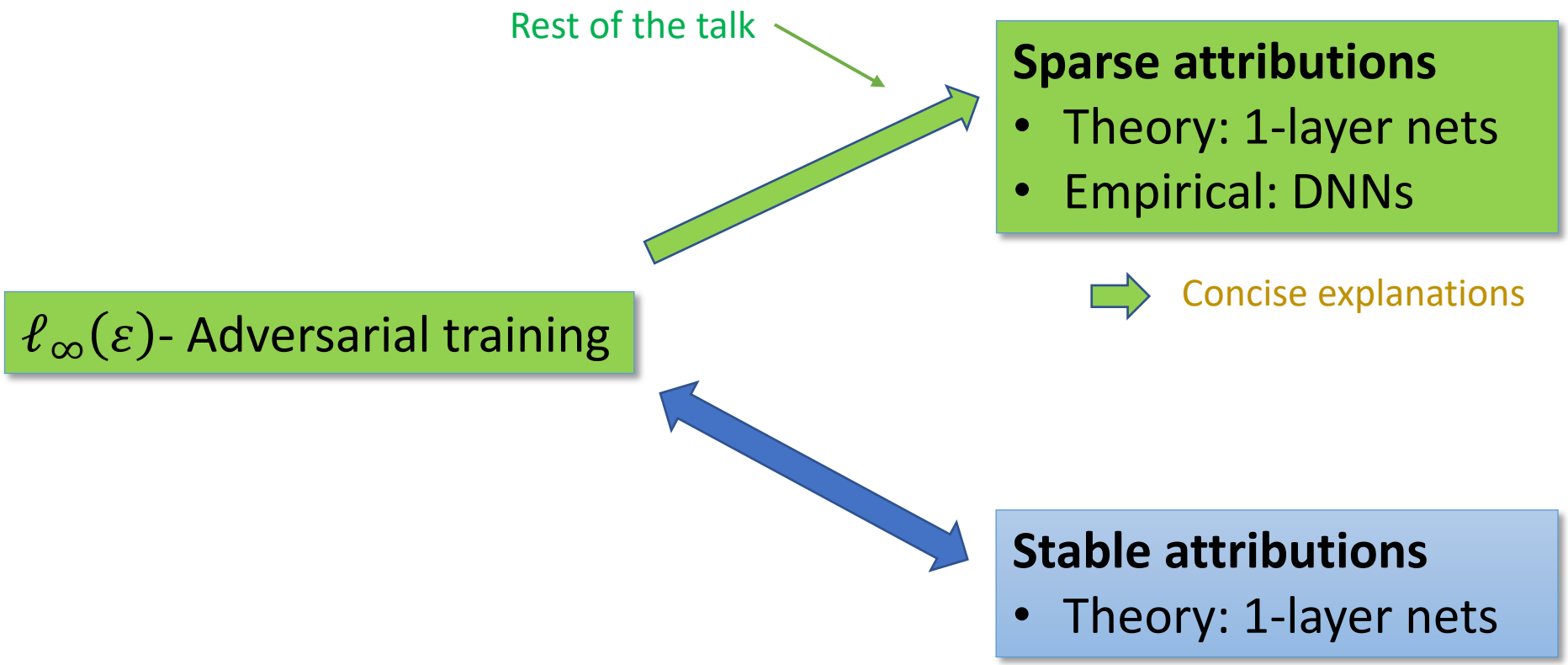
i.e., train on "worst-case" inputs $\boldsymbol{x}+\delta$,
where δ is constrained by adversary's "budget" ε

Our results:



"Robust models have better explanations"

Our results:



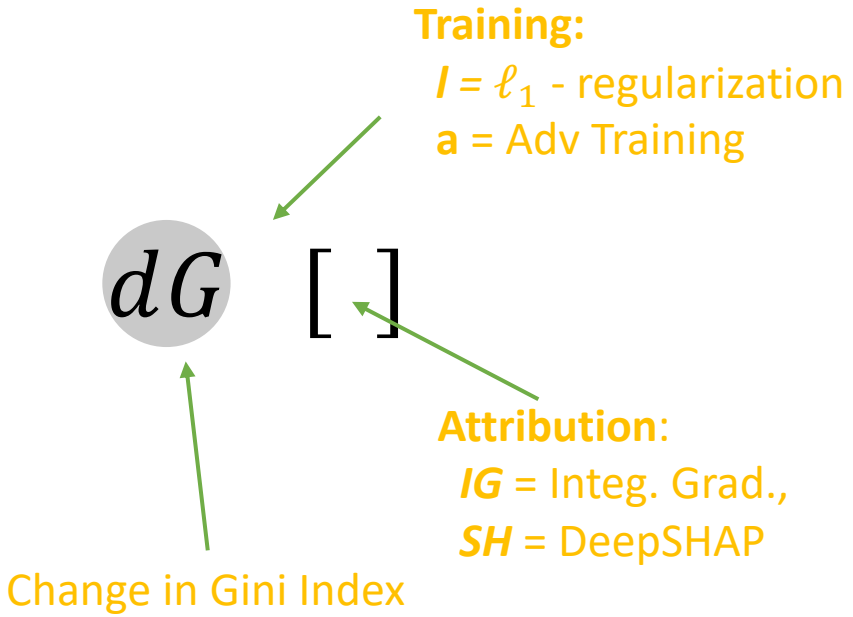
"Robust models have better explanations"

But what about ℓ_1-regularization?

ℓ_1 −regularization	$\ell_\infty(\varepsilon)$-adversarial training
• →sparse weights $\boldsymbol{w}$	• → sparse attributions $A^F(\boldsymbol{x})$
• Agnostic of output $F(\boldsymbol{x};\ \boldsymbol{w})$	• Explicitly depends on $F(\boldsymbol{x};\boldsymbol{w})$

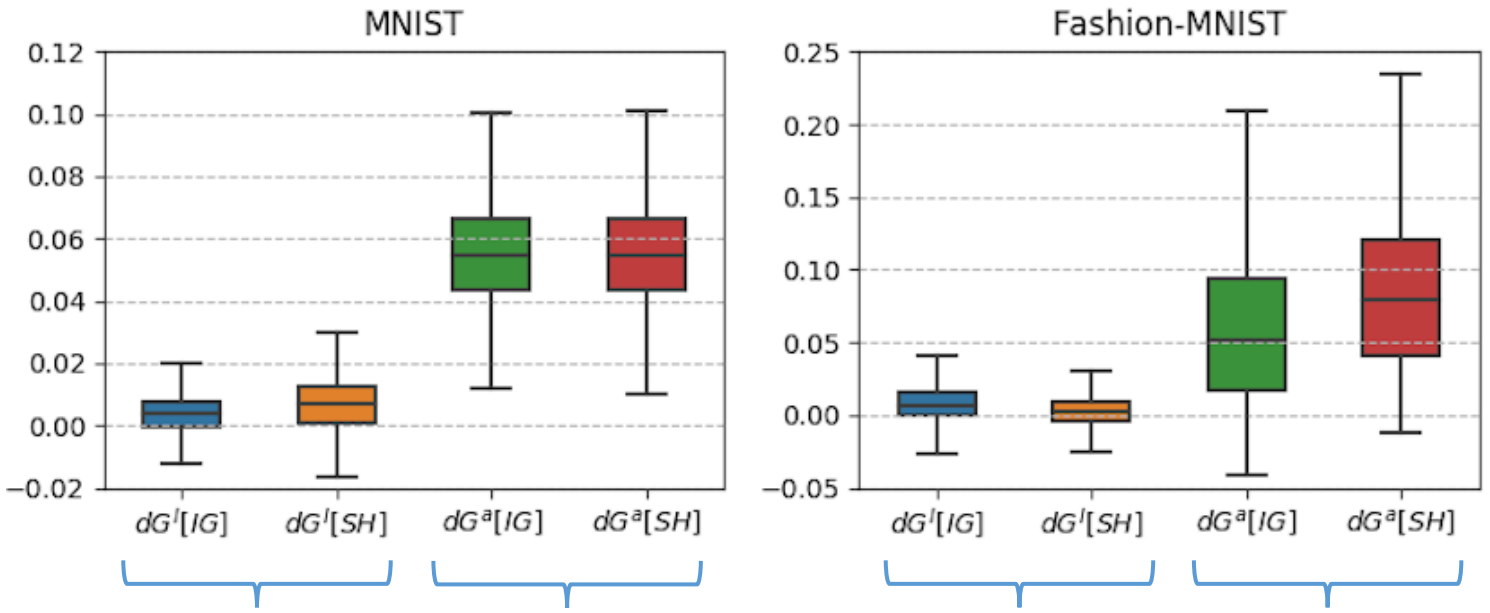
Results on MNIST, Fashion-MNIST

Gini Index $G(|\boldsymbol{v}|)$ measures sparseness of an attribution vector $\boldsymbol{v}$



Results on MNIST, Fashion-MNIST

Point-wise change in Gini-index



ℓ_1- reg $\lambda = 0.01$ | adversarial $\varepsilon = 0.3$

ℓ_1- reg $\lambda = 0.01$ | adversarial $\varepsilon = 0.1$

Gini Index $G(|\boldsymbol{v}|)$ measures sparseness of an attribution vector $\boldsymbol{v}$

Training:
l = ℓ_1 - regularization
a = Adv Training

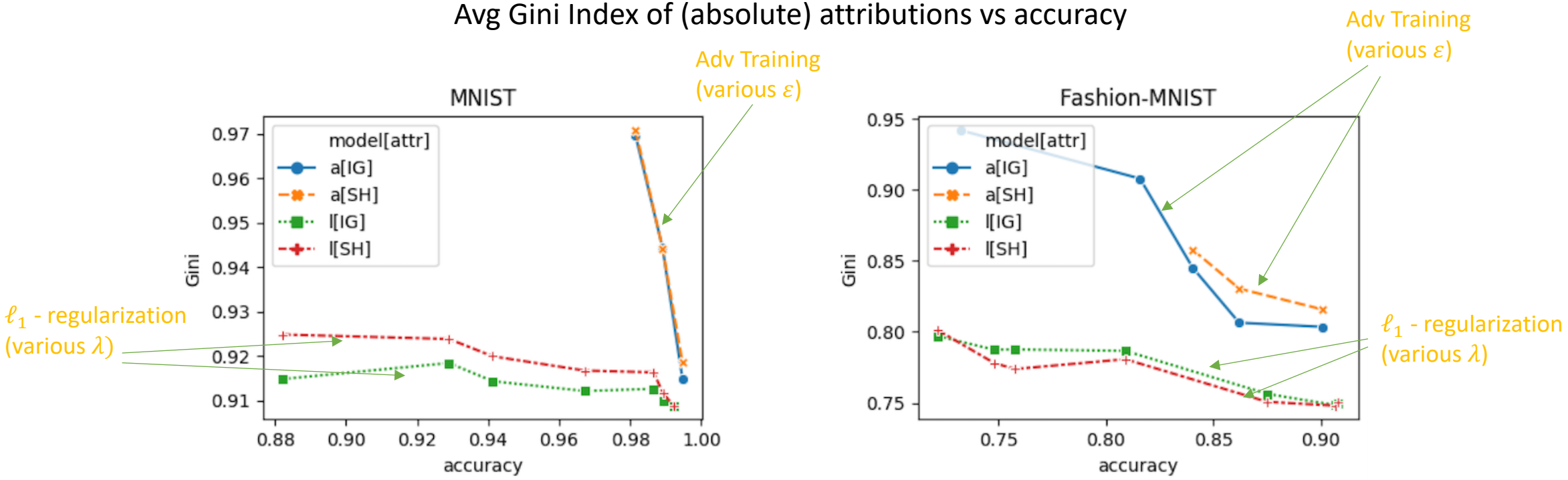
dG [ ]

Attribution:
IG = Integ. Grad.,
SH = DeepSHAP

Change in Gini Index

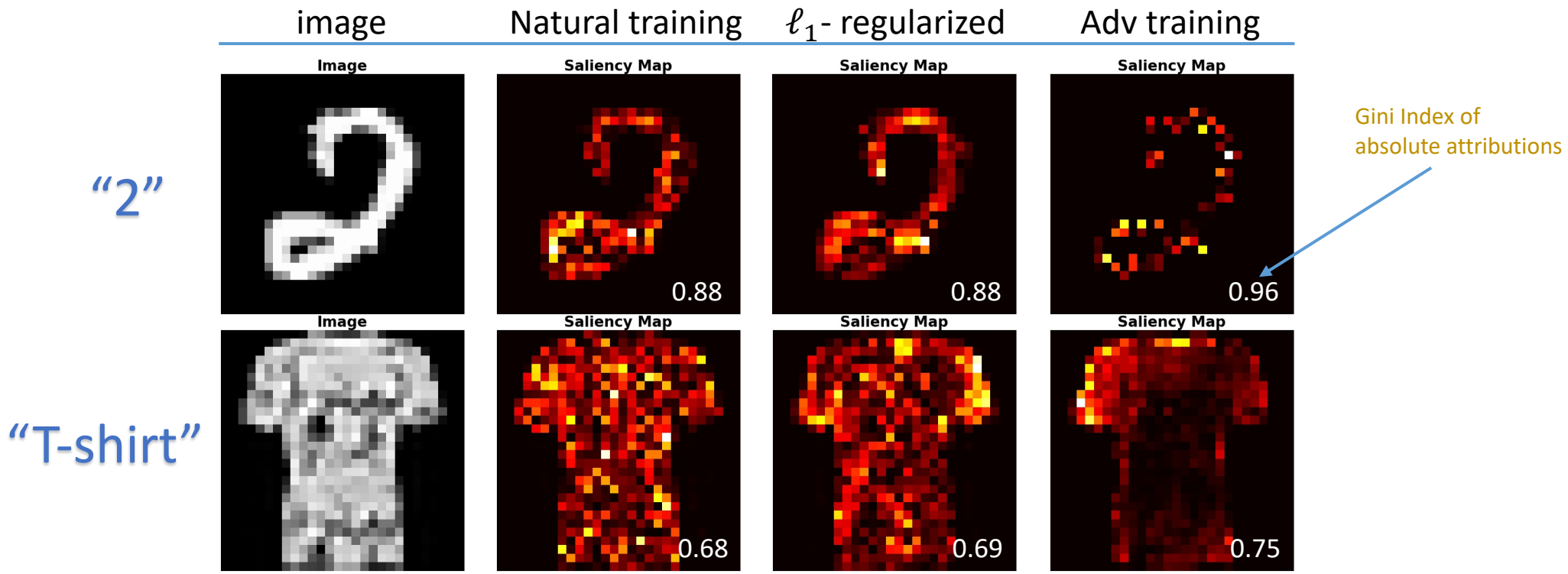
Results on MNIST, Fashion-MNIST

Avg Gini Index of (absolute) attributions vs accuracy



$\ell_\infty(\varepsilon)$-adversarial training >> ℓ_1-regularization

Attribution sparseness on images



Thank you!

pchalasani@xaipient.com