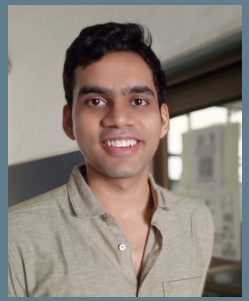
Use Perturbations when Learning from Explanations

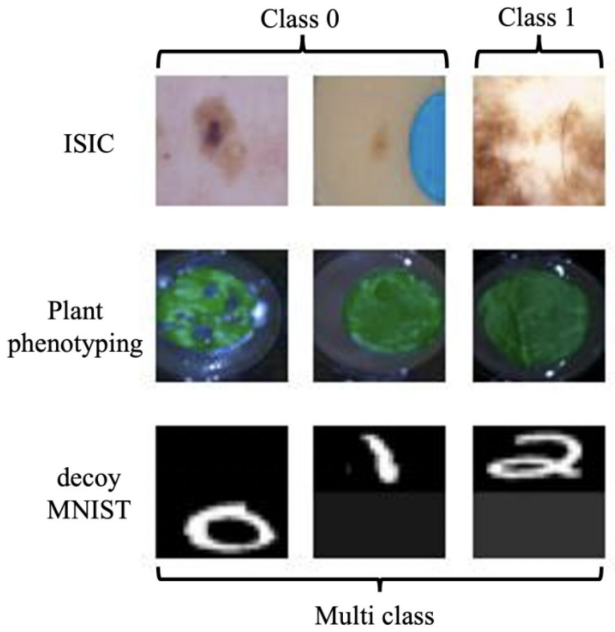
Juyeon Heo*, Vihari Piratla*, Matthew Wicker, Adrian Weller

NeurIPS 2023

Motivation

1. Remove dependency on the patch for classifying the skin lesion as non-cancerous
2. Do not depend on the background for classifying the leaf as healthy or not
3. Decoy-MNIST: Ignore the label-revealing half and only rely on digit half to label the image.

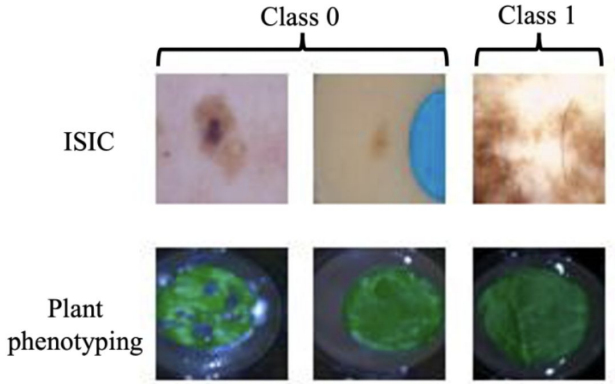


Motivation (continued…)

Motivation: It is surprisingly hard to alter ML model's prediction behavior

Popular solutions:

1. Collect massively diverse data to negate such spurious correlation
2. Collect some data from each group and impose an invariance constraint across environments.



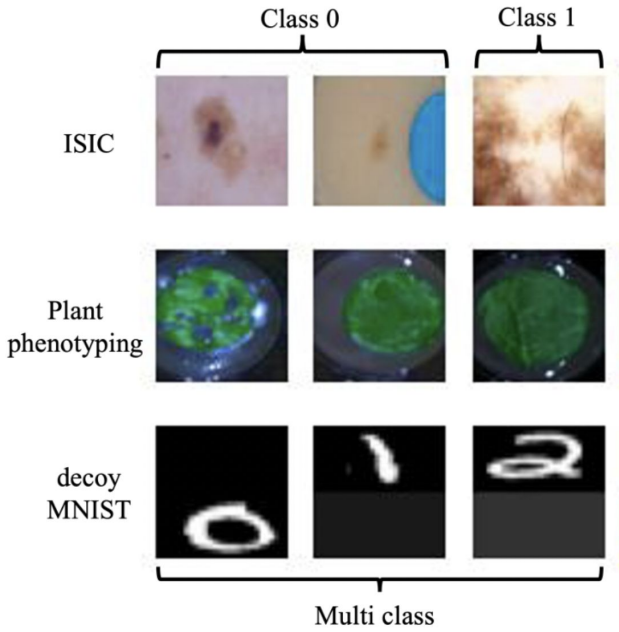
Collecting more data is very expensive (especially in healthcare)
Invariance-like methods still suffer when the minority group is not sufficiently large

Our problem and objective

Objective: Let user provide richer annotation and use it to learn better aligned models

Problem

1. Process inputs along with saliency map highlighting nuisance features
2. Learn a model that ignores the irrelevant regions highlighted by the explanation.



Background

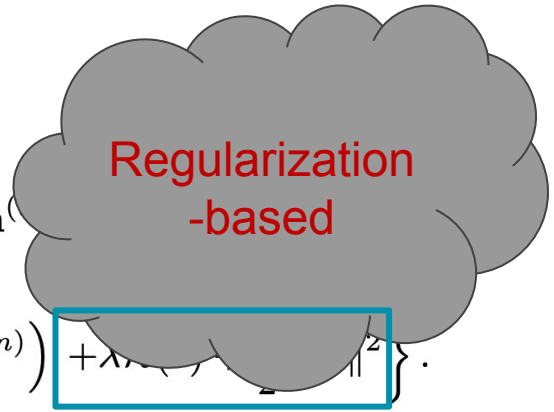
Standard approach: train a model such that saliency of masked region is near zero.

An explanation algorithm (E) to assign importance scores to input features: IS(x), which is then regularized with an R(θ) term such that irrelevant features are not regarded as important.

$$IS(\mathbf{x}) \triangleq E(\mathbf{x}, f(\mathbf{x};\theta)).$$

$$\mathcal{R}(\theta) \triangleq \sum_{n=1}^{N} \| IS(\mathbf{x}^{(n)}) \odot \mathbf{m}^{(}$$

$$\theta^* = \arg\min_{\theta} \left\{ \sum_{n} \ell\left(f(\mathbf{x}^{(n)};\theta), y^{(n)}\right) \right.$$



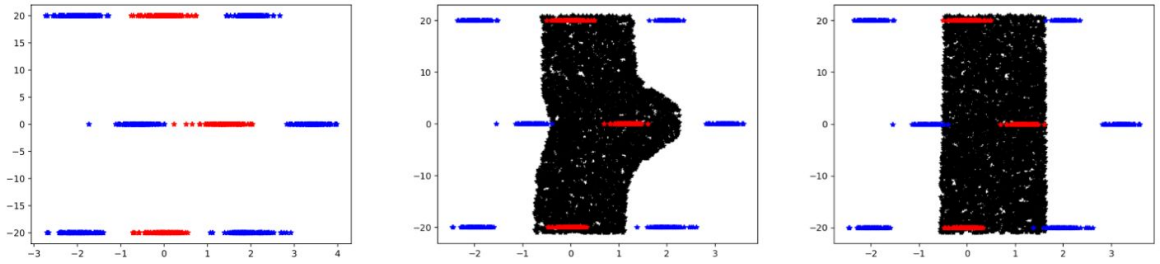
Ross et.al. 2017, Rieger et.al. 2020, Shao et.al. 2021

Background

Explanation Algorithm	Citation
Gradient explanation	Grad-Reg (Ross et.al. 2017)
Contextual decomposition based explanation (Singh et al. (2018))	CDEP (Rieger et.al. 2020)
Influence functions	Shao et al. (2021)
Concept Explanations	Stammer et.al. (2021)
LIME	Schramowski, P. (2020)

Our findings

Regularization-based although popular are not effective



(a) Toy data (b) Grad-Reg with β=0 (c) Grad-Reg with β=1

Figure 1: Illustration of the uneasy relationship between Grad-Reg and smoothing strength. (b) The decision boundary is nearly vertical (zero gradient wrt to nuisance y-axis value) for all training points and yet varies as a function of y value when Grad-Reg fitted using $\beta = 0$. (c) Grad-Reg requires strong model smoothing ($\beta = 1$) in order to translate local insensitivity to global robustness to x-coordinate. (d) IBP-Ex fits vertical pair of lines without any model smoothing.

Analysis for why robustness-based better than regularization-based

$$f \sim GP(0, K); k(x, \tilde{x}) = \exp(-\sum_i \frac{1}{2} \frac{(x_i - \tilde{x}_i)^2}{\theta_i^2});$$

$$\theta_i^{-2} \sim \mathcal{G}(\alpha, \beta); \text{i.e. } \mathbb{E}[\theta_i^{-2}] = \alpha/\beta$$

Theorem 1 (Grad-Reg). *We infer a regression function f from a GP prior as described above with the additional supervision of $[\partial f(\mathbf{x})/\partial x_2]|_{\mathbf{x}^{(i)}} = 0, \quad \forall i \in [1, N]$. Then the function value deviations to perturbations on irrelevant feature are lower bounded by a value proportional to the perturbation strength δ as shown below.*

$$f(\mathbf{x} + [0, \delta]^T) - f(\mathbf{x}) \geq \frac{2\delta\alpha}{\beta} \Theta(x_1^2 x_2^6 + \delta x_1^2 x_2^5) \tag{5}$$

α/β must be set to a small value, i.e. theta large to limit function deviations.

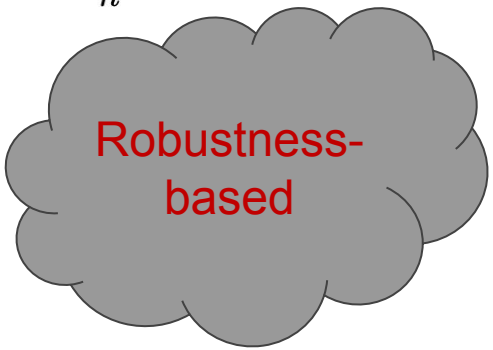
I.e. the function must be smooth for the regularization of local explanations to apply globally.

Robustness instead of regularization

Robustness to perturbation drawn from human specification

Our methodology is built on the interpretation of the provided mask as a specification of a low-dimensional manifold from which input perturbations are drawn.

$$\theta^* = \arg\min_{\theta} \sum_{n} \left\{ \ell\left(f(\mathbf{x}^{(n)};\theta), y^{(n)}\right) + \alpha \max_{\boldsymbol{\epsilon}:\|\boldsymbol{\epsilon}\|_\infty \leq \kappa} \ell\left(f(\mathbf{x}^{(n)} + (\boldsymbol{\epsilon} \odot \mathbf{m}^{(n)});\theta), y^{(n)}\right) \right\}$$



Robustness instead of regularization (continued…)

– – –

$$\theta^* = \arg\min_\theta \sum_n \left\{ \ell\left(f(\mathbf{x}^{(n)};\theta), y^{(n)}\right) + \alpha \max_{\boldsymbol{\epsilon}:\|\boldsymbol{\epsilon}\|_\infty \le \kappa} \ell\left(f(\mathbf{x}^{(n)} + (\boldsymbol{\epsilon} \odot \mathbf{m}^{(n)});\theta), y^{(n)}\right) \right\}$$

	How the inner maximization is solved
Avg-Ex	Approximate max with MC average
PGD-Ex	Solve inner maximization using PGD (Madry et.al. 2017)
IBP-Ex	Minimize an upper certifiable bound on the worst perturbation (possible for smallish networks) (Mirman et.al. 2018, Gowan et.al. 2018)

Promise of robustness-based methods

Define C as the maximum distance of any point in the domain to the closest point covered by the defense, i.e. C is the strength of the defense method

Theorem 2. *When we use a robustness algorithm to regularize the network, the fitted function has the following property.*

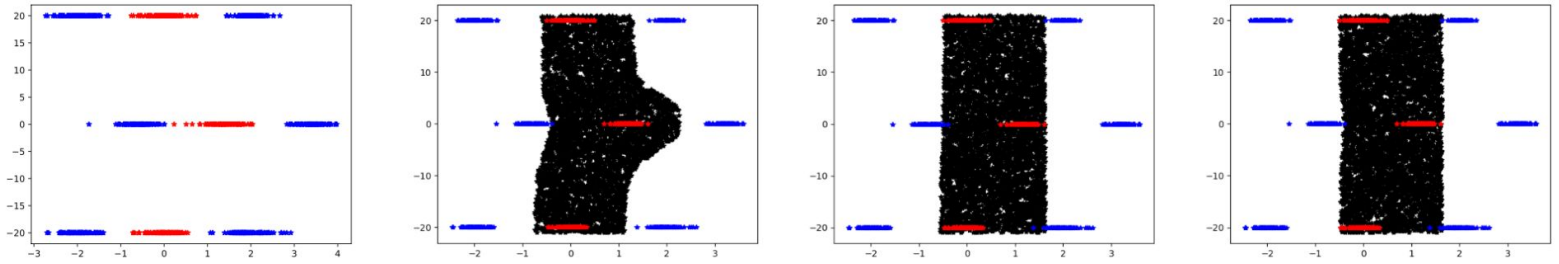
$$|f(\mathbf{x} + [0, \delta]^T) - f(\mathbf{x})| \leq 2C\frac{\alpha}{\beta}\delta_{max} f_{max}. \tag{6}$$

δ_{max} and f_{max} are maximum values of Δx_2 and $f(\mathbf{x})$ in the input domain ($\mathcal{X}$) respectively.

Can bound the deviation if the defense is decent (C is small) without having to smooth the function

Our findings

Regularization-based although popular are not effective



(a) Toy data (b) Grad-Reg with β=0 (c) Grad-Reg with β=1 (d) IBP-Ex with β=0

Figure 1: Illustration of the uneasy relationship between Grad-Reg and smoothing strength. (b) The decision boundary is nearly vertical (zero gradient wrt to nuisance y-axis value) for all training points and yet varies as a function of y value when Grad-Reg fitted using $\beta = 0$. (c) Grad-Reg requires strong model smoothing ($\beta = 1$) in order to translate local insensitivity to global robustness to x-coordinate. (d) IBP-Ex fits vertical pair of lines without any model smoothing.

Robustness-based have curse of dimensionality

– – –

$$\theta^* = \arg\min_{\theta} \sum_{n} \left\{ \ell\left(f(\mathbf{x}^{(n)};\theta), y^{(n)}\right) + \alpha \max_{\boldsymbol{\epsilon}:\|\boldsymbol{\epsilon}\|_\infty \leq \kappa} \ell\left(f(\mathbf{x}^{(n)} + (\boldsymbol{\epsilon} \odot \mathbf{m}^{(n)});\theta), y^{(n)}\right) \right\}$$

Inner maximization quickly gets harder to solve as the input dimensionality increases.

Proposition 1. *Consider a regression task with $D+1$-dimensional inputs $\mathbf{x}$ where the first D dimensions are irrelevant, and assume they are $x_d = y, d \in [1, D]$ while $x_{D+1} \sim \mathcal{N}(y, 1/K)$. The MAP estimate of linear regression parameters $f(\mathbf{x}) = \sum_{d=1}^{D+1} w_d x_d$ when fitted using Avg-Ex are as follows: $w_d = 1/(D+K), \quad d \in [1, D]$ and $w_{D+1} = K/(K+D)$.*

Our Work

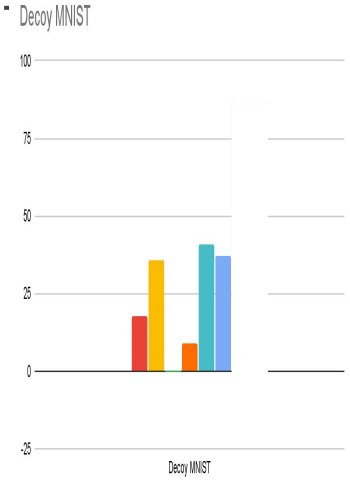
Could be advantageous to combine robustness and regularization-based methods for complementary strengths

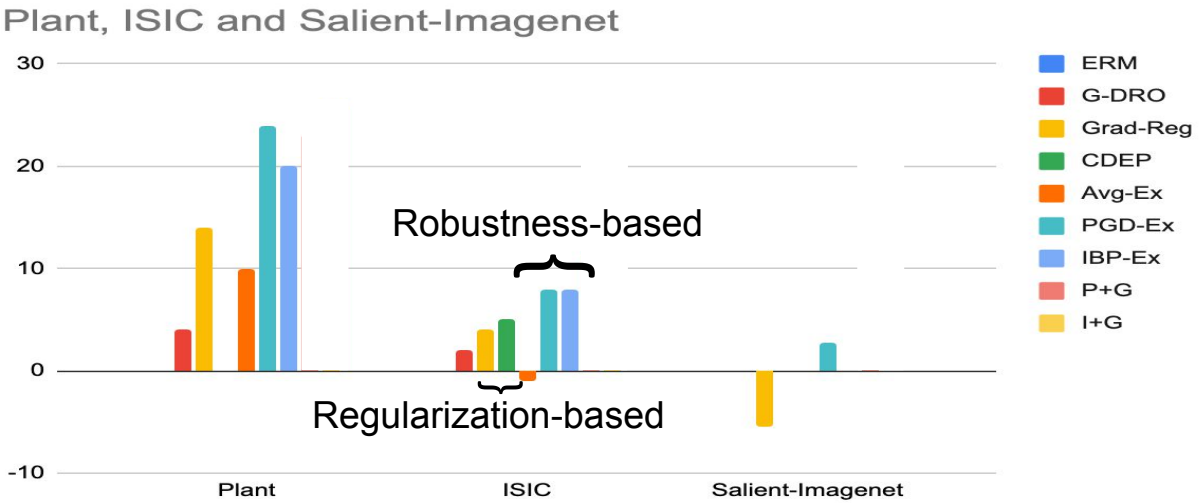
We did a systematic study of

(a) Regularization-based: Grad-Reg, CDEP
(b) Robustness-based: PGD-Ex, IBP-Ex
(c) Their combination: IBP-Ex + Grad-Reg, PGD-Ex + Grad-Reg

for the problem of effective learning from human-specified explanations.

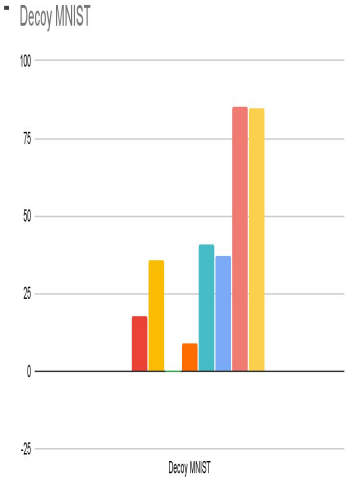
Results

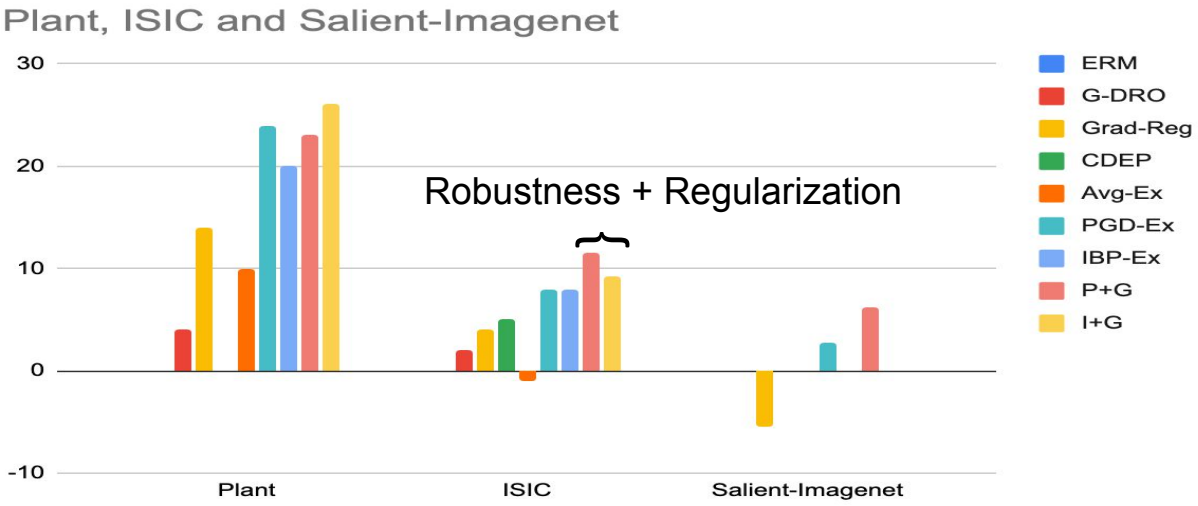




Robustness-based methods are better

Results





Regularization-based

Regularization + robustness is even better as expected

Results

Dataset→	Decoy-MNIST	Plant	ISIC	S-Imagenet
Method↓	Robust Accuracy			
ERM	10.5 ± 5.4	54.8 ± 1.3	55.9 ± 2.3	87.5
G-DRO	28.1 ± 0.1	58.0 ± 4.6	58.5 ± 10.7	-
Grad-Reg	46.2 ± 1.1	68.2 ± 1.4	60.2 ± 7.4	82.2
CDEP	10.0 ± 0.7	54.2 ± 24.7	60.9 ± 3.0	-
Avg-Ex	19.5 ± 1.4	64.5 ± 0.3	55.2 ± 6.6	-
PGD-Ex	51.4 ± 0.3	78.5 ± 0.3	$\mathbf{64.4 \pm 4.3}$	90.2
IBP-Ex	47.6 ± 2.0	73.8 ± 1.7	64.2 ± 1.2	-

Robustness-based methods are better

Results

Regularization + robustness is even better as expected

ERM	10.5 ± 5.4	54.8 ± 1.3	55.9 ± 2.3	87.5
G-DRO	28.1 ± 0.1	58.0 ± 4.6	58.5 ± 10.7	-
Grad-Reg	46.2 ± 1.1	68.2 ± 1.4	60.2 ± 7.4	82.2
CDEP	10.0 ± 0.7	54.2 ± 24.7	60.9 ± 3.0	-
Avg-Ex	19.5 ± 1.4	64.5 ± 0.3	55.2 ± 6.6	-
PGD-Ex	51.4 ± 0.3	78.5 ± 0.3	$\mathbf{64.4 \pm 4.3}$	90.2
IBP-Ex	47.6 ± 2.0	73.8 ± 1.7	64.2 ± 1.2	-
P+G	$\mathbf{95.8 \pm 0.4}$	76.7 ± 2.8	$\mathbf{67.5 \pm 1.1}$	**93.8**
I+G	$\mathbf{95.0 \pm 0.6}$	$\mathbf{80.1 \pm 0.3}$	$\mathbf{65.2 \pm 1.8}$	-

Takeaways

- We presented intuition and theoretical analysis for why regularization-based methods are not suited for supervising with human explanations.
- We studied robustness-based method for supervising with human-explanations, which is surprisingly not studied before.
- Our systematic study and analysis showed advantage in combining robustness and regularization-based methods for effective supervision.

Future and Limitations

Reducing human effort.

(a) Assistance in providing human explanations
(b) Partial specifications
(c) Automated discovery of regions