Robust Attribution Regularization

By Jiefeng Chen Xi Wu Vaibhav Rastogi Yingyu Liang Somesh Jha

02/14/2019

Presenter: Zijie Pan

https://qdata.github.io/deep2Read/

Motivation

- Deep Learning is treated as black box, which is too much to understand or interpret

- Robust attribution plays important fundamental role for humans in classification tasks, but only recently, draw attentions to ML area

- Lack of attention makes DL vulnerable to adversarial examples:

    - Brittle predictions: model robustness
    - Brittle attributions: Explanation robustness

Proposed Solution

- Add robust attribution regularization term in training

- RAR aims to regularize the training so the resulting model will have robust attributions that are not substantially changed under minimal input perturbations.

Preliminary Concept

- Attribution:

Compare the DNN output F(x) to what its output would have been if the input feature were xi were not active (replace by some information-less baseline value bi)

Formula：

$$A_i^F(x;b) = F(x) - F(x[x_i = b_i])$$

Preliminary Concept

- Axiom of Attribution:

  - Completeness or Additivity: Sum of feature attribution equals to F(x)

  - Sensitivity: For non-zero feature and $F(x) \neq 0$,attribution of that feature is not zero

  - Implementation Variance: When two neural network compute the same mathematical function, regardless how differently they are implemented, the attributions for all features should be the same

Preliminary Concept

- Axiom of Attribution Cont.:

    - Linearity: compose two NN,H = aF+bG, indicates attributions are the weighted sum

    - Symmetry-Preserving: For any input x where the values for two symmetric features (interchange them does not change the function mathematically) are the same, the attributions should be the same.

      Symmetric features Ex: F(x) = min(1,x1+x2)

Related Work

- Based on proof from economic side knowledge(Friedman, Eric J et al..):
  Path Integrated Gradient method to calculate attribution satisfies all axioms except last one.

  Path function: x=g(α). Infinite number of possible paths available
  The attribution of the feature at dimension i can be calculated as:

$$A_i^{F,\Pi}(x) = \int_0^1 \partial_i F(g(\alpha)) \frac{\partial g_i(\alpha)}{\partial \alpha} d\alpha.$$

- Paper by Sundararajan et. al states that attribution using the Integrated Gradient along the **straight line** from the origin to x is the unique Path Method that also satisfies the last axiom.

uniformly scaling: gi(α) = αxi, so the derivative term equals xi and the function simplifies to:

$$A_i^F(x) = x_i \int_0^1 \partial_i F(\alpha x) d\alpha$$

- The paper uses the general formulation

Claim / Target Task

- Using IG method to quantify attributions
- Robust Attribution Regularization:

$$\underset{\theta}{\text{minimize}} \quad \underset{(\boldsymbol{x},y)\sim P}{\mathbb{E}}[\rho(\boldsymbol{x}, y; \theta)]$$

$$\text{where } \rho(\boldsymbol{x}, y; \theta) = \ell(\boldsymbol{x}, y; \theta) + \lambda \max_{\boldsymbol{x}' \in N(\boldsymbol{x}, \varepsilon)} s(\mathrm{IG}_{\boldsymbol{h}}^{\ell_y}(\boldsymbol{x}, \boldsymbol{x}'; r))$$

- P : data distribution
- θ: Model parameter set
- λ : regularization parameter
- x: input
- x’: perturbed input
- IG:Give the attribution of features respect to the changes of loss value (apply to intermediate layer h)
- s: size function

Formula Insight

- RAR gives principled generalizations of objective designed for robust predictions in both uncertainty set model and distributional robustness model

- Uncertainty set model:
  - (Madry et al) λ =1 and size function is Sum() and L_∞-Norm bounded perturbation
    ρ(x,y;θ) = $\max_{\boldsymbol{x}' \in N(\boldsymbol{x},\varepsilon)} \ell(\boldsymbol{x}', y; \theta).$
  - Input gradient regularization $\rho(\boldsymbol{x}, y; \theta) = \ell(\boldsymbol{x}, y; \theta) + \lambda \|\nabla_{\boldsymbol{x}} \ell(\boldsymbol{x}, y; \theta)\|_q^q.$
  - Regularization by attribution of the loss output:
    $$\rho(\boldsymbol{x}, y; \theta) = \ell(\boldsymbol{x}, y; \theta) + \max_{\boldsymbol{x}' \in N(\boldsymbol{x},\varepsilon)} \{|\ell_y(\boldsymbol{x}') - \ell_y(\boldsymbol{x})|\}$$

- Distributional Robustness Model
  - Wasserstein prediction robustness

$$\underset{\theta}{\text{minimize}} \quad \left\{ \underset{P}{\mathbb{E}}[\ell(P;\theta)] + \lambda \sup_{Q;M \in \prod(P,Q)} \left\{ \underset{M=(Z,Z')}{\mathbb{E}} \left[d_{\mathrm{IG}}(Z, Z') - \gamma c(Z, Z') \right] \right\} \right\}$$

Formula Insight Cont.

- for 1-layer neural networks, RAR naturally degenerates to max-margin training.

Implementation

- IG-NORM: Size function is L1-Norm

$$\underset{\theta}{\text{minimize}} \quad \mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim P}\left[\ell(\boldsymbol{x},y;\theta) + \lambda \max_{\boldsymbol{x}'\in N(\boldsymbol{x},\varepsilon)} \|\,\mathrm{IG}^{\ell_y}(\boldsymbol{x},\boldsymbol{x}')\|_1\right]$$

- IG-SUM-NORM: s(·) = sum(·) + β*L1-Norm(·)

$$\underset{\theta}{\text{minimize}} \quad \mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim P}\left[\max_{\boldsymbol{x}'\in N(\boldsymbol{x},\varepsilon)} \left\{\ell(\boldsymbol{x}',y;\theta) + \beta\|\,\mathrm{IG}^{\ell_y}(\boldsymbol{x},\boldsymbol{x}')\|_1\right\}\right]$$

- SGD Training
- Attack: PDG attack

Data Summary

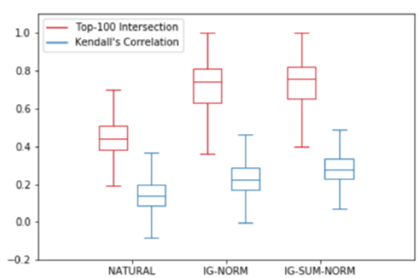
- MNIST, Fashion-MNIST, GTSRB, Flower
- Evaluation: Accuracy+Kendall’s tau rank order correlation+Top-k intersection

Experimental Results

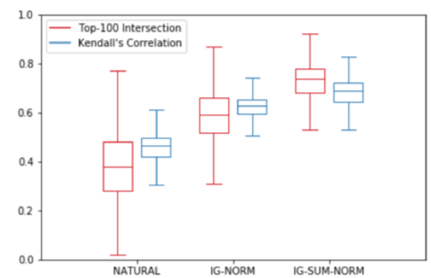
Dataset	Approach	Nat Acc.	Adv Acc.	TopK Inter.	Rank Corr.
MNIST	NATURAL	99.17%	0.00%	46.61%	0.1758
	Madry et al.	98.40%	92.47%	62.56%	0.2422
	IG-NORM	98.74%	81.43%	71.36%	0.2841
	IG-SUM-NORM	98.34%	88.17%	**72.45%**	**0.3111**
Fashion-MNIST	NATURAL	90.86%	0.01%	39.01%	0.4610
	Madry et al.	85.73%	73.01%	46.12%	0.6251
	IG-NORM	85.13%	65.95%	59.22%	0.6171
	IG-SUM-NORM	85.44%	70.26%	**72.08%**	**0.6747**
GTSRB	NATURAL	98.57%	21.05%	54.16%	0.6790
	Madry et al.	97.59%	83.24%	68.85%	0.7520
	IG-NORM	97.02%	75.24%	**74.81%**	0.7555
	IG-SUM-NORM	95.68%	77.12%	74.04%	**0.7684**
Flower	NATURAL	86.76%	0.00%	8.12%	0.4978
	Madry et al.	83.82%	41.91%	55.87%	0.7784
	IG-NORM	85.29%	24.26%	64.68%	0.7591
	IG-SUM-NORM	82.35%	47.06%	**66.33%**	**0.7974**

Experimental Results

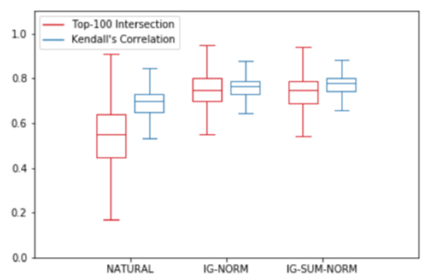
- Compared with naturally trained model, RAR only sacrifice small drops on testing accuracy. (Right thing to do, not learning spurious relationships)
- But gives robust predictions and robust attribution



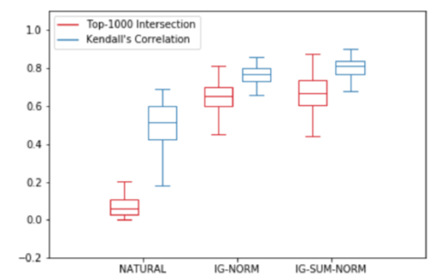
(a) MNIST



(b) Fashion-MNIST



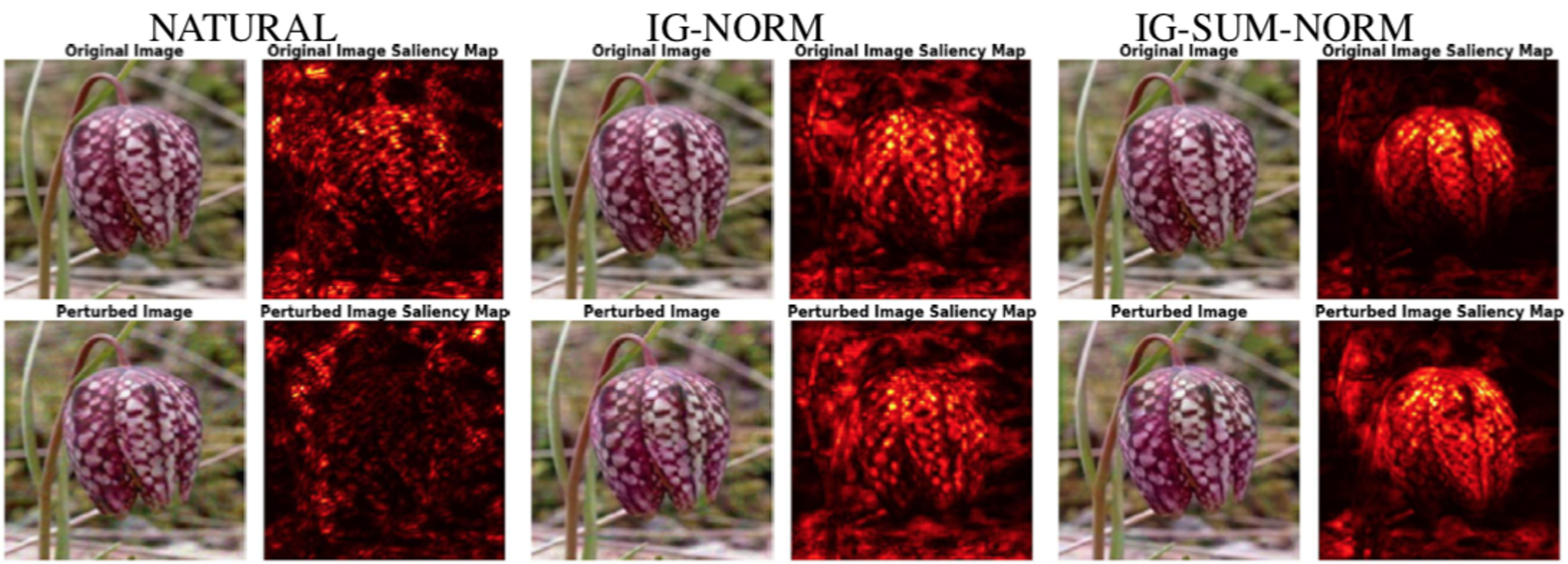
(c) GTSRB



(d) Flower

Experimental Results

- Very Interesting: RAR leads to much human aligned attributions
  We can explicitly see the highlighted attributions are flower-shaped.



NATURAL	IG-NORM	IG-SUM-NORM
Top-1000 Intersection: 0.1%	Top-1000 Intersection: 58.8%	Top-1000 Intersection: 60.1%
Kendall's Correlation: 0.2607	Kendall's Correlation: 0.6736	Kendall's Correlation: 0.6951

Project Idea

Model to learn robust attributions and connect to explainable model.

Using robust attribution training as feature extractions and feed into looks like model