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
Challenges

• Blackbox: not too much understanding/interpretation
Challenges

- Blackbox: not too much understanding/interpretation

- Vulnerable to adversaries
Interpretable Machine Learning

• Attribution task: Given a model and an input
Interpretable Machine Learning

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

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

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**: $\text{Attr}(a*f_1 + b*f_2) = a*\text{Attr}(f_1) + b*\text{Attr}(f_2)$

- **Completeness**: $\sum(\text{Attr}) = f(\text{input}) - f(\text{baseline})$

- **Symmetry Preservation**: Symmetric variables with identical values get equal attributions.
Integrated Gradient: Example Results

Original image

Top label: stopwatch
Score: 0.998507

Integrated gradients

Original image

Top label: jackfruit
Score: 0.99591

Integrated gradients

Original image

Top label: school bus
Score: 0.997033

Integrated gradients
Observation 1: Attribution is Fragile.

Interpretation of Neural Networks is Fragile.
Observation 1: Attribution is Fragile

Interpretation of Neural Networks is Fragile.
Observation 1: Attribution is Fragile

Interpretation of Neural Networks is Fragile.
Observation 1: Attribution is Fragile

Interpretation of Neural Networks is Fragile.
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

\[ \min_\theta \ E[l(x, y; \theta) + \lambda * \text{RAR}] \]
Robust Attribution Regularization

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

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

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

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

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

$$RAR = \max_{x' \in \Delta(x)} s(IG(x, x'))$$
Robust Attribution Regularization

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

\[ \min_\theta \mathbb{E}[l(x, y; \theta) + \lambda \times RAR] \]

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

\[
\min_\theta \ E[l(x, y; \theta) + \lambda \cdot RAR]
\]

\[
RAR = \max_{x' \in \Delta(x)} s(IG(x, x'))
\]

• Two instantiations:

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

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

- Top-1000 Intersection
- Kendall's Correlation

Comparison across:
- NATURAL
- IG-NORM
- IG-SUM-NORM
Result on MINST dataset
Result on GTSRB dataset
Connection to Robust Prediction

• RAR

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

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

• If \( \lambda = 1 \) and \( s(\cdot) = \text{sum}(\cdot) \), then RAR becomes the Adversarial Training objective for robust prediction

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

simply by the Completeness of IG

Towards Deep Learning Models Resistant to Adversarial Attacks.
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 \( s(\cdot) = \text{sum}(\cdot) \) coincide, and both reduce to soft max-margin training.
Connection to Robust Prediction

• RAR

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

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

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

\[
\min_\theta \mathbb{E}[l(x, y; \theta) + \lambda'\|\nabla_x l(x, y; \theta)\|_q^q]
\]

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!