BIRS Workshop on Interpretability in AI

Causal Perspectives in Explaining Neural Network Models

4 May 2022

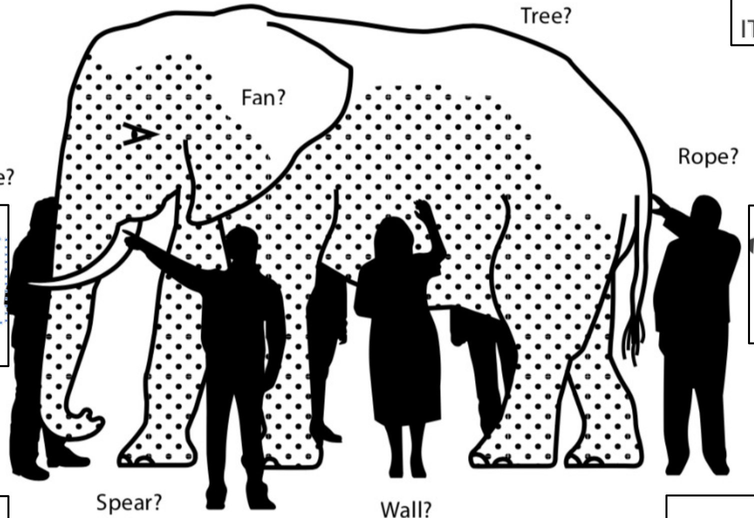
Vineeth N Balasubramanian
Department of Computer Science and Engineering/Artificial Intelligence
Indian Institute of Technology, Hyderabad

आई आई टी हैदराबाद
IIT Hyderabad

The Elephant in the Room

What is it really?

Functional: For those who use it



IT & Operations

How do I monitor and debug this model?

Business Owner

Can I trust our AI decisions?

Internal Audit, Regulators

Are these AI system decisions fair?

Data Scientists

Is this the best model that can be built?

Customer Support

How do I answer this customer complaint?

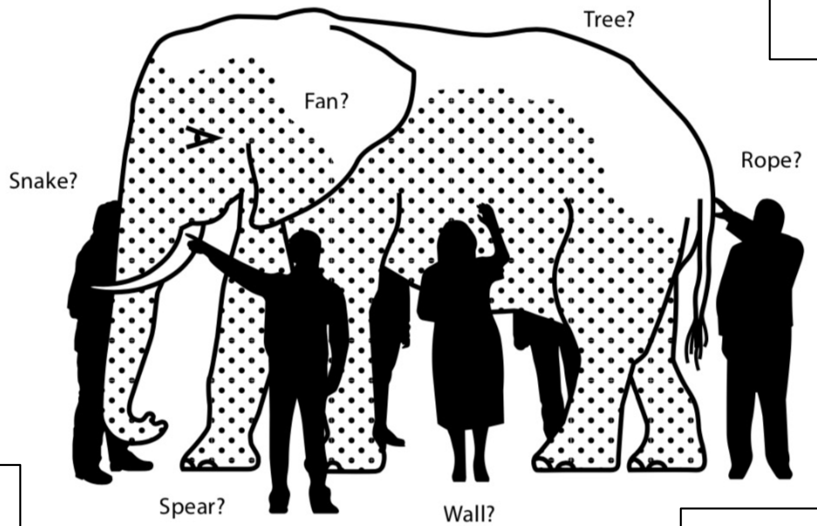
The Elephant in the Room

What is it really?

Technical: For those who build it

Global (vs) Local explanations

Post-hoc explainable (vs) Intrinsically interpretable



Model-agnostic (vs) Model-specific approaches

Transparency (vs) Reasoning

Attributions (vs) Actionable Explanations

Causal (vs) Correlational associations

Feature-level (vs) Latent Concept-level Explanations

Viewing XAI from Different Perspectives

Our Efforts

Non-Causal

Post-hoc Explainability

* [WACV 2018] GradCAM++: Generic method for visual explanations for CNN models
* [IEEE Trans on Biometrics 2021] Canonical saliency maps for face recognition/processing models
* [AISTATS 2022] Submodular ensembles of attribution methods

Causal

* [ICML 2019] Causal attributions in neural networks

Explainability in Deep Learning

Intrinsic Interpretability

Non-Causal

* [CVPR 2022] Ante-hoc explainability via concepts
* [CVPR 2022] Transferring concepts in knowledge distillation tasks

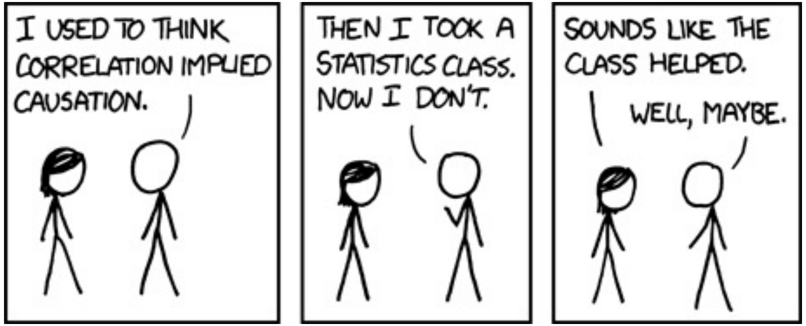
Causal

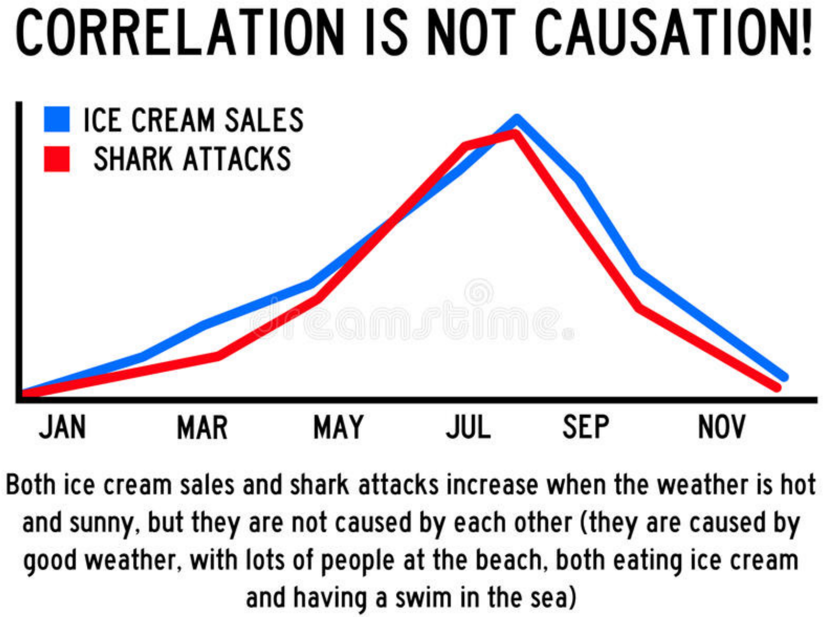
* [AAAI 2022] Causally disentangled representations
* [CVPR'W 2021] Dataset for causal representation learning
* [WACV 2022] Mitigating bias through causal perspectives
* [arXiv 2022] Causal regularizers

Complementarity of explanations and robustness [AAAI 2021, NeurIPS 2021]

Causation vs Correlation in XAI

- Is feature correlation to output a true indicator of explainability?
- Or do we need to find causal relationships in the analyzed data-output pairs?





Causality in Machine Learning

The Three Layer Causal Hierarchy

Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y\|x)$	Seeing	What is? How would seeing X change my belief in Y?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y\|do(x), z)$	Doing Intervening	What if? What if I do X?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x\|x', y')$	Imagining, Retrospection	Why? Was it X that caused Y? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smoking the past 2 years?

Figure 1: The Causal Hierarchy. Questions at level i can only be answered if information from level i or higher is available.

Judea Pearl, The Seven Tools of Causal Inference with Reflections on Machine Learning, 2018
Judea Pearl, The Book of Why: The New Science of Cause and Effect, 2018

Causal Perspectives in XAI

Our Recent Efforts

Given a trained NN model, what causal input-output attributions did it learn?

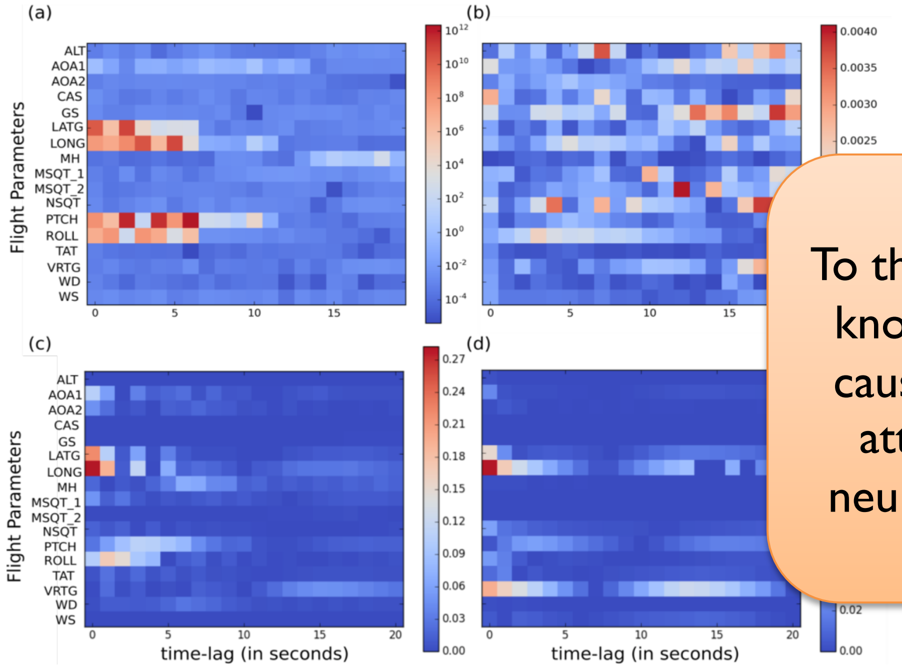
Given known causal domain priors of input-output relationships, can we make a trained NN model learn and maintain these causal relationships?

Causal Attributions in Neural Networks
ICML 2019

Causal Regularization with Domain Priors
arXiv, 2022 (under review)

Causal Attributions in Neural Networks

ICML 2019



To the best of our knowledge, first causal effort for attribution in neural networks

Joint work with:

Aditya Chattopadhyay

Piyushi Manupriya

Anirban Sarkar

आई आई टी हैदराबाद
IIT Hyderabad

Causal Attributions of Neural Network Models

What does this mean?

- **Attribution:** Effect of an input feature on prediction function's output
  - Inherently a causal question!
- Existing attribution methods
  - Gradient-based
    - "How much would perturbing a particular input affect the output?" Not a causal analysis
  - Using surrogate models (or interpretable regressors)
    - Correlation-based again

Causal Attributions of Neural Network Models

What does this mean?

- **Our objective:** What are the causal attributions learned by a trained neural network model?
  - To the best of our knowledge, first such effort
- Assume a setting that is often valid
  - Input dimensions are causally independent of each other (they can be jointly caused by a latent confounder)
- Show how this can be done with feedforward networks as well as RNNs

Structural Causal Model

Structural Causal Model

$$(X, U, f, P_u))$$

Endogenous variables

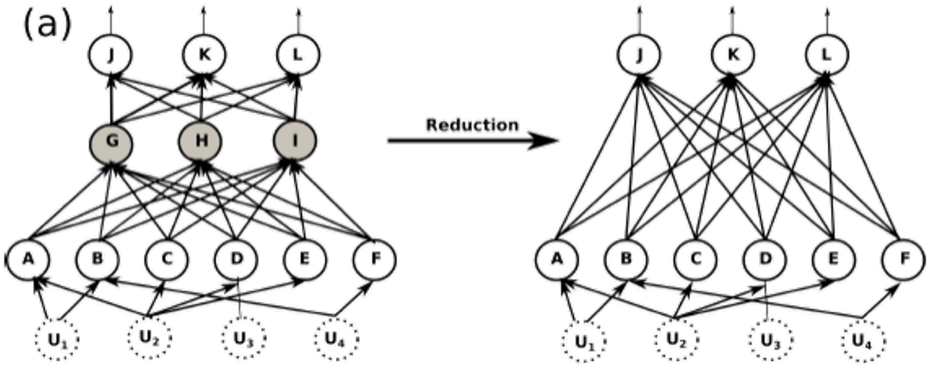
Exogenous variables

Causal Functions

Distribution of U

Neural Network as a SCM

Feedforward neural network

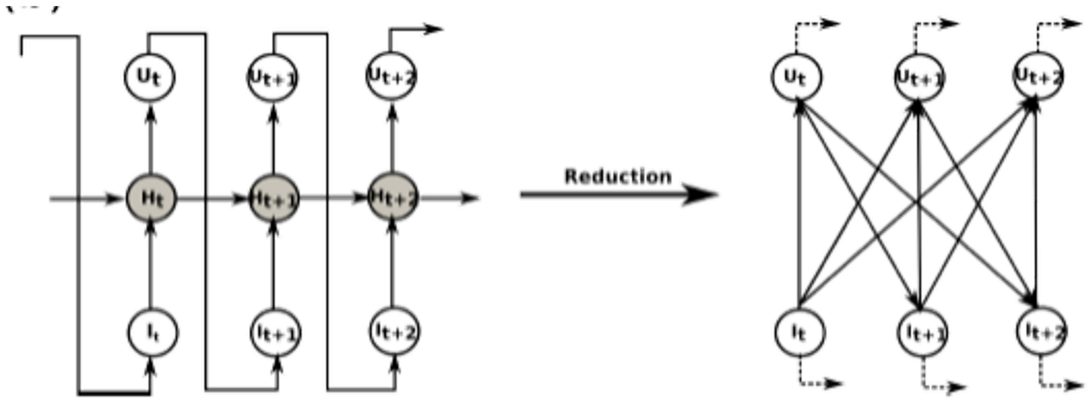


$$M([l_1, l_2, ..., l_n], U, [f_1, f_2, ...f_n], P_U) \qquad M'([l_1, l_n], U, f', P_U)$$

- l_i – neurons in layer I
- f_i – corresponding causal functions

Neural Network as a SCM

Recurrent neural network



Gradient-Based Attribution

- Gradient-based and Perturbation-based attribution methods – special cases of Individual Causal Effect

$$ICE^{y}_{do(x_i=\alpha)} = y_{x_i=\alpha}(u) - y(u)$$

- Setting α to $u_i + \epsilon$
- Such methods are sensitive and cannot give global attributions

Causal Attribution

ACE: Average Causal Effect

For binary variables:

$$\mathbb{E}[y|do(x=1)] - \mathbb{E}[y|do(x=0)]$$

For continuous variables:

$$ACE^{y}_{do(x_i=\alpha)} = \mathbb{E}[y|do(x_i=\alpha)] - baseline_{x_i}$$

where baseline is defined as:

$$\mathbb{E}_{x_i}[\mathbb{E}_y[y|do(x_i=\alpha)]]$$

the average ACE across all x_i

Interventional expectation: Non-trivial to compute

Computing ACE

$$\mathbb{E}[y|do(x_i = \alpha)] = \int_y yp(y|do(x_i = \alpha))dy$$

Interventional expectation: Non-trivial to compute

Let:

$$y = f'_y(x_1, x_2, ..., x_k)$$

$$\mu_j = \mathbb{E}[x_j|do(x_i = \alpha)] \forall x_j \in l_1$$

$$\mu = [\mu_1, \mu_2, ..., \mu_k]^T$$

Consider the Taylor-series expansion:

$$f'_y(l_1) \approx f'_y(\mu) + \nabla^T f'_y(\mu)(l_1 - \mu) + \frac{1}{2}(l_1 - \mu)^T \nabla^2 f'_y(\mu)(l_1 - \mu)$$

Marginalizing over all other input neurons:

$$\mathbb{E}[f'_y(l_1)|do(x_i = \alpha)] \approx f'_y(\mu) + \frac{1}{2} Tr(\nabla^2 f'_y(\mu) \mathbb{E}[(l_1 - \mu)(l_1 - \mu)^T |do(x_i = \alpha)])$$

Computing ACE

$$\mathbb{E}[y|do(x_i = \alpha)] = \int_y yp(y|do(x_i = \alpha))dy \longrightarrow \mathbb{E}[f'_y(l_1)|do(x_i = \alpha)] \approx f'_y(\mu) + \frac{1}{2}Tr(\nabla^2 f'_y(\mu)\,\mathbb{E}[(l_1 - \mu)(l_1 - \mu)^T|do(x_i = \alpha)])$$

- Intervened input neuron is **d-separated** from other input neurons; what does this give us?
- Given an intervention on a particular variable, the probability distribution of all other input neurons doesn't change, i.e. for $x_j \neq x_i$

$$P(x_j|do(x_i = \alpha)) = P(x_j)$$

- Interventional means and covariances of non-intervened neurons same as observational means and covariances

Can be pre-computed

Causal Regressors

Computing the baseline

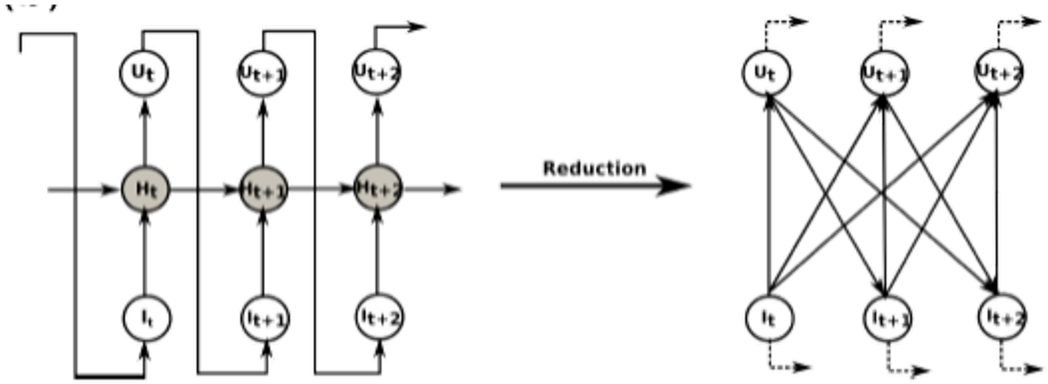
$$ACE^{y}_{do(x_i=\alpha)} = \mathbb{E}[y|do(x_i = \alpha)] - baseline_{x_i}$$

where baseline is defined as:

$$\mathbb{E}_{x_i}[\mathbb{E}_y[y|do(x_i = \alpha)]]$$

We use causal regressors (Bayesian regression) to obtain baseline using different intervention values, α, from its range

What about RNNs?



Depends on a particular RNN architecture.

Where output does not feed into input, same idea can be used.

Scaling to Large Data

Computation of ACE requires Hessian:

$$\mathbb{E}[f'_y(l_1)|do(x_i=\alpha)] \approx f'_y(\mu) + \frac{1}{2}Tr(\nabla^2 f'_y(\mu)\mathbb{E}[(l_1-\mu)(l_1-\mu)^T|do(x_i=\alpha)])$$

However, we only need

$$\sum_{i=1}^{k}\sum_{j=1}^{k}\nabla^2 f'_y(\mu)_{ij}Cov(x_i,x_j|do(x_l=\alpha))$$

To this end, consider eigendecomposition of covariance matrix:

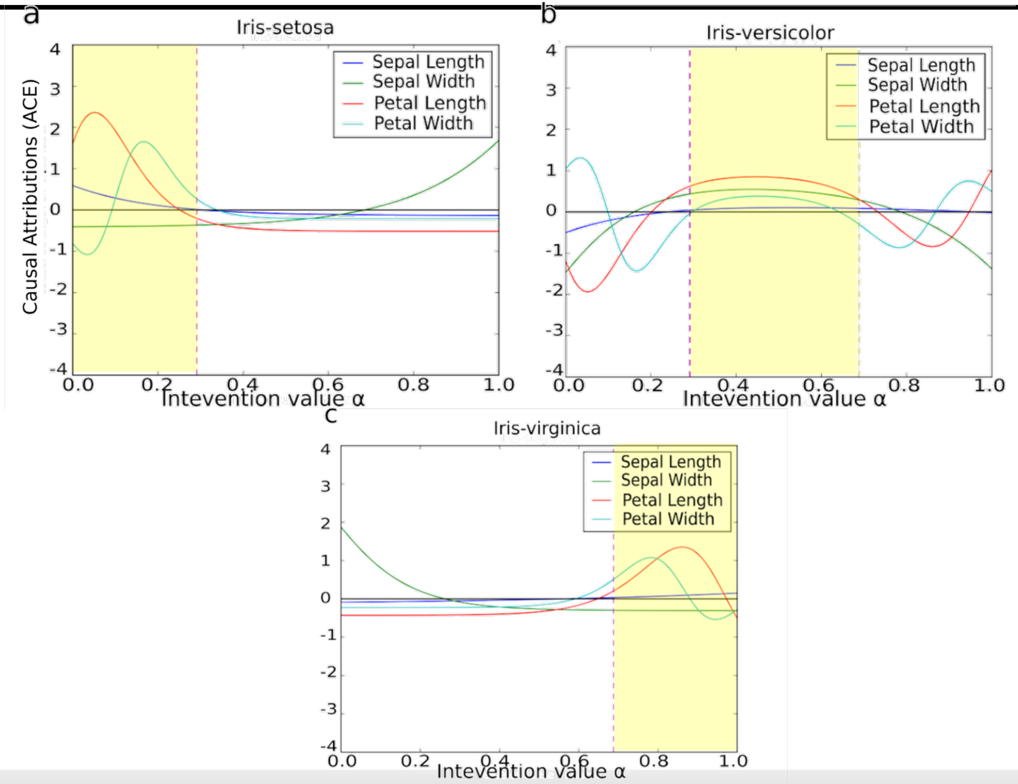
$$Cov(\mathbf{x},\mathbf{x}|do(x_l=\alpha)) = \sum_{r=1}^{\kappa}\lambda_r e_r e_r^T$$

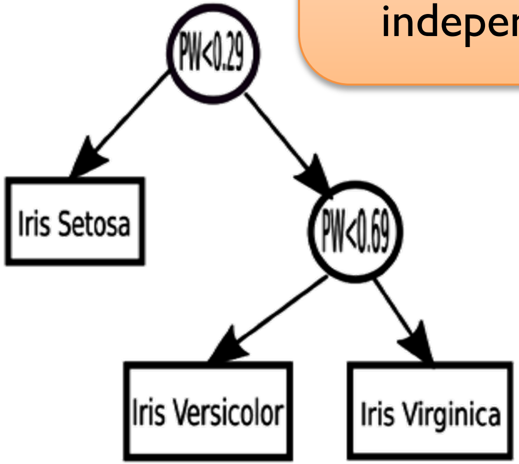
Let $v_r = \lambda^{1/2} e_r$

$$\lim_{\epsilon\to 0}\frac{1}{\epsilon^2}\Big(f'_y(\mu-\epsilon v_r)+f'_y(\mu+\epsilon v_r)-2f'_y(\mu)\Big) = v_r^T\nabla^2 f'_y(\mu)v_r$$

Results

Iris Dataset



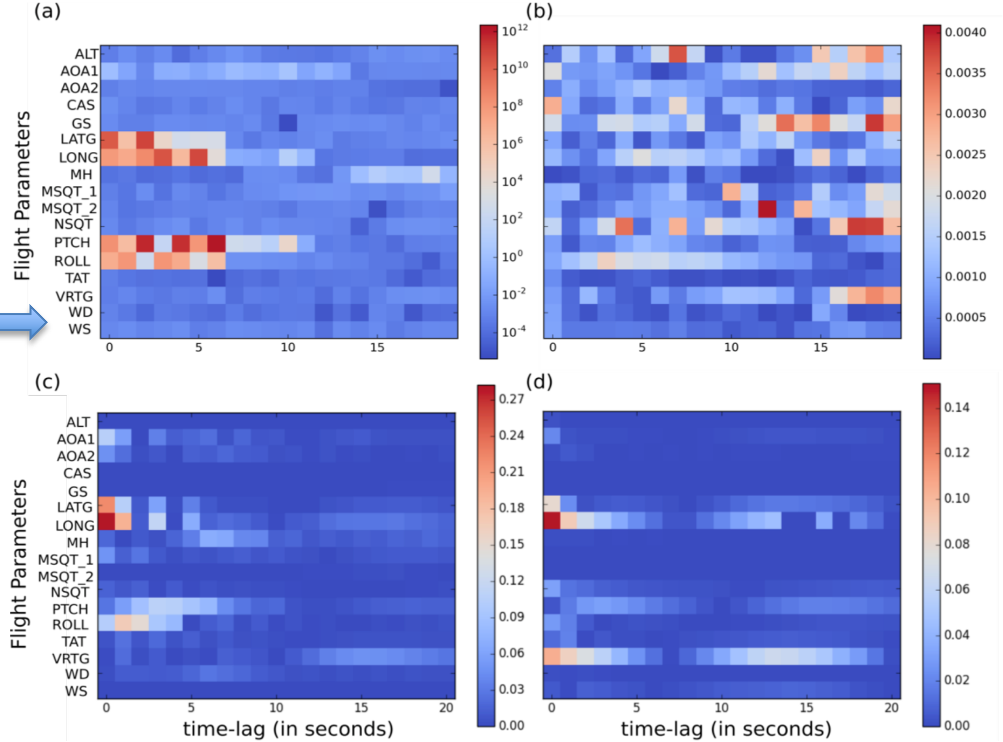


ACE values match a decision tree learned independently!

Results

Aircraft Data (NASA Dashlink Dataset)

FDR report: "….*due to slippery runway, the pilot could not apply timely brakes, resulting in a steep acceleration in the airplane post-touchdown…*"



Axioms of Attribution

- Completeness
- Sensitivity
- Implementation Invariance
- Symmetry Preservation
- Input Invariance

Proposed method satisfies all important axioms (almost)

Completeness: *For any input x, the sum of the feature attributions equals $F(x)$:* $F(x) = \sum_i A_i^F(x)$

Sensitivity: *If x has only one non-zero feature and $F(x) \neq 0$, then the attribution to that feature should be non-zero.*

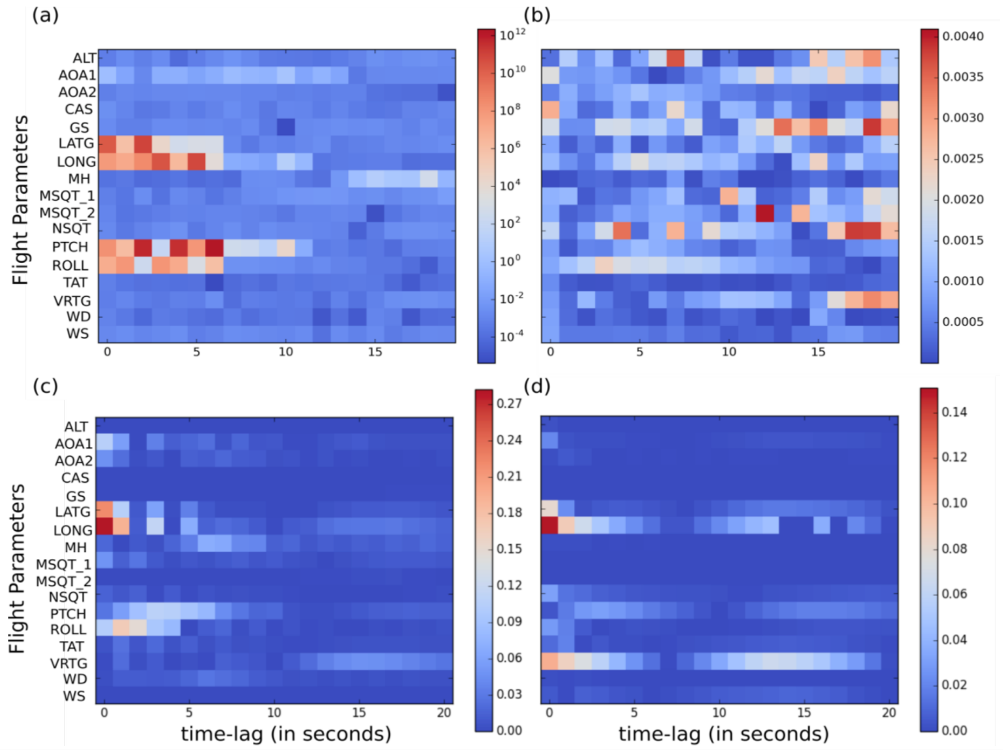
Implementation Invariance: *When two neural networks compute the same mathematical function $F(x)$, regardless of how differently they are implemented, the attributions to all features should always be identical.*

Symmetry-Preserving: *For any input x where the values of two symmetric features are the same, their attributions should be identical as well.*

Gradient-based methods violate Axiom 2;
DeepLIFT and LRP violate Axiom 3

Sundararajan et al, ICML 2017; Kindermans et al, 2017

More Details

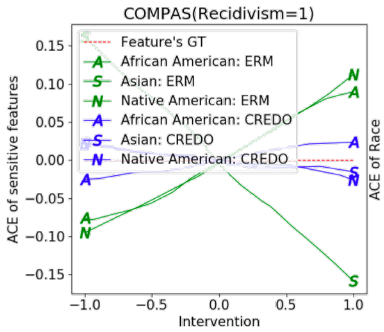


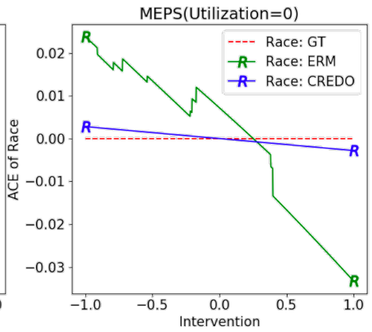
arXiv:
https://arxiv.org/abs/1902.02302

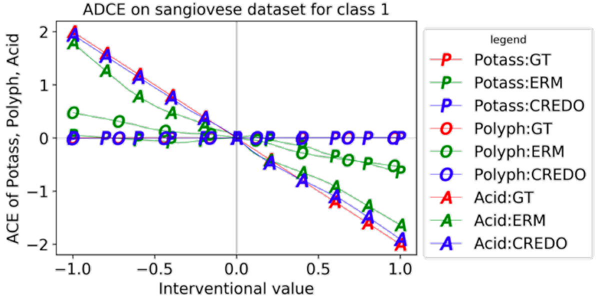
Code:
https://github.com/Piyushi-0/ACE

Causal Regularization with Domain Priors

arXiv 2022
(under review)







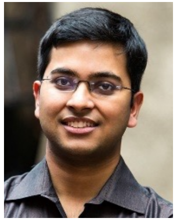
To the best of our knowledge, first effort to integrate causal knowledge for attribution in neural networks

Joint work with:

Gowtham Reddy A

Sai Srinivas K

Amit Sharma

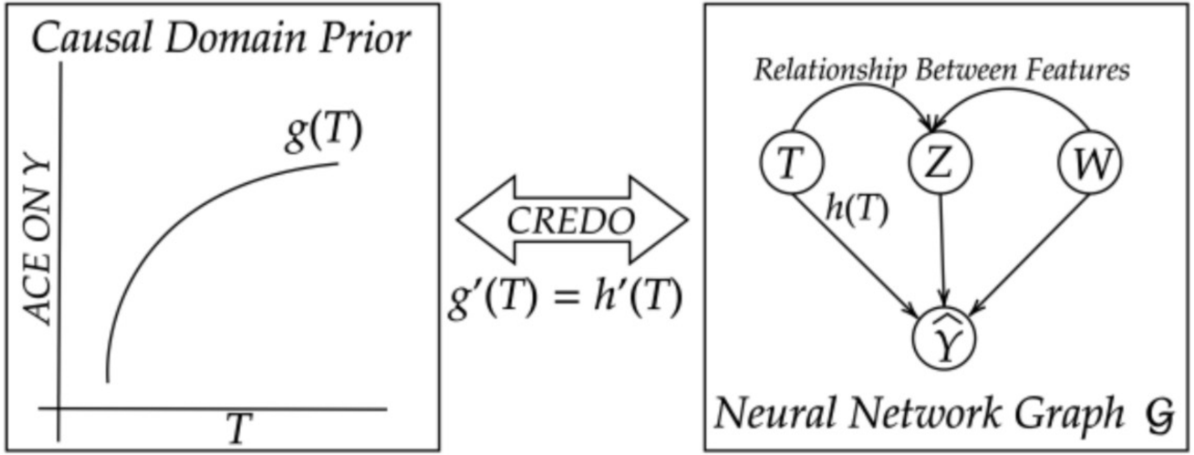


Microsoft Research

आई आई टी हैदराबाद
IIT Hyderabad

Key Idea

- Match causal effects learned by a neural network to effects we want it to learn



CREDO: Causal **RE**gularization with **DO**main Priors

Causal Graph and Effects

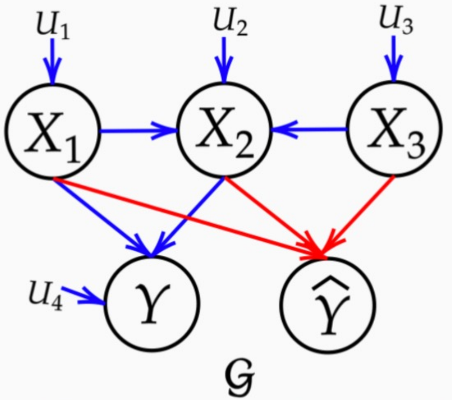


Figure 5: Causal graph $\mathcal{G}$ representing input features X_1, X_2, X_3, true output Y and NN output $\hat{Y}$ *(Blue arrows = true causal relationships, Red arrows = relationships learned by traditional NN (without CREDO).*

We regularize for three kinds of causal effect in NN models:

- Controlled direct effect
- Natural direct effect
- Total causal effect

Matching Controlled Direct Effect

Let $Y_{x=\alpha} := Y|do(x=\alpha)$

Definition

(Controlled Direct Effect in NN). Controlled Direct Effect $(NN-CDE)$ measures the causal effect of treatment T at an intervention t (i.e., $do(T=t)$) on $\hat{Y}$ when all parents of $\hat{Y}$ except T $(PA^{\hat{Y}})$ are intervened to pre-defined control values α. *Average Controlled Direct Effect* $(NN-ACDE)$ is defined as: $NN-ACDE^{\hat{Y}}_{t,PA^{\hat{Y}}=\alpha} :=$ $\mathbb{E}_U[\hat{Y}_{t,PA^{\hat{Y}}=\alpha}] - \mathbb{E}_U[\hat{Y}_{t^*,PA^{\hat{Y}}=\alpha}] = \hat{Y}_{t,PA^{\hat{Y}}=\alpha} - \hat{Y}_{t^*,PA^{\hat{Y}}=\alpha}$.

$$NN-ACDE^{\hat{Y}}_t := \mathbb{E}_{PA^{\hat{Y}}}[\hat{Y}_{t,PA^{\hat{Y}}}] - \mathbb{E}_{PA^{\hat{Y}}}[\hat{Y}_{t^*,PA^{\hat{Y}}}]$$

Regularizing for Controlled Direct Effect

Proposition

(ACDE Identifiability in Neural Networks) For a neural network with output $\hat{Y}$, the ACDE of a feature T at t on $\hat{Y}$ is identifiable and given by $ACDE_t^{\hat{Y}} = \mathbb{E}_{PA^{\hat{Y}}}\left[\hat{Y}|t, PA^{\hat{Y}}\right] - \mathbb{E}_{PA^{\hat{Y}}}\left[\hat{Y}|t^*, PA^{\hat{Y}}\right]$.

Proposition

(ACDE Regularization in Neural Networks) The n^{th} partial derivative of $ACDE$ of T at t on $\hat{Y}$ is equal to the expected value of n^{th} partial derivative of $\hat{Y}$ w.r.t. T at t, that is: $\frac{\partial^n ACDE_t^{\hat{Y}}}{\partial t^n} = \mathbb{E}_{PA^{\hat{Y}}}\left[\frac{\partial^n[\hat{Y}(t, PA^{\hat{Y}})]}{\partial t^n}\right]$.

Our Regularizer

$$\hat{\theta} = \arg\min_{\theta} ERM + \lambda \frac{1}{N} \sum_{j=1}^{N} \max\{0, \|\nabla_j f \odot M - \delta G^j\|_1 - \epsilon\}$$

where $\nabla_j f$ is the $C \times d$ Jacobian of f w.r.t. x^j; M is a $C \times d$ binary matrix that acts as an indicator of features for which prior knowledge is available; $\odot$ represents the element-wise (Hadamard) product; N is the size of training data; and ϵ is a hyperparameter to allow a margin of error.

Algorithm 1 CREDO Regularizer

Result: Regularizers for ACDE, ANDE, ATCE in f.

Input: $\mathcal{D} = \{(x^j, y^j)\}_{j=1}^{N}$, $y^j \in \{0, 1, \dots, C\}$, $x^j \sim X^j$; $\mathbb{Q} = \{i | \exists\, g_i^c \ for\ some\ c\}$; $\mathbb{G} = \{g_i^c | g_i^c$ is prior for i^{th} feature w.r.t. class c$\}$; $\mathbb{F} = \{f^1, \dots, f^K\}$ is the set of structural equations of the underlying causal model s.t f^i describes Z^i; ϵ is a hyperparameter

Initialize: $j = 1, \delta G^j = \mathbb{0}_{c \times d} \forall j = 1, ..., N, M = \mathbb{0}_{c \times d}$

while $j \leq N$ **do**

  foreach $i \in \mathbb{Q}$ **do**

    foreach $g_i^c \in \mathbb{G}$ **do**

      $\delta G^j[c, i] = \nabla g_i^c|_{x_i^j}$; $M[c, i] = 1$

      case *1: regularizing ACDE* **do**

        $\nabla_j f[c, i] = \frac{\partial \hat{Y}}{\partial x_i}|_{x^j}$

      case *2: regularizing ANDE* **do**

        /* causal graph is known */

        $t = x_i$

        $\nabla_j f[c, i] = \frac{\partial \hat{Y}}{\partial x_i}|_{(t^j, z_{i*}^j, w^j)}$

      case *3: regularizing ATCE* **do**

        /* causal graph is known */

        $\nabla_j f[c, i] = \left[\frac{d\hat{Y}}{dx_i} + \sum_{l=1}^{K} \frac{\partial \hat{Y}}{\partial Z^l} \frac{df^l}{dx_i}\right]|_{x^j}$

    end

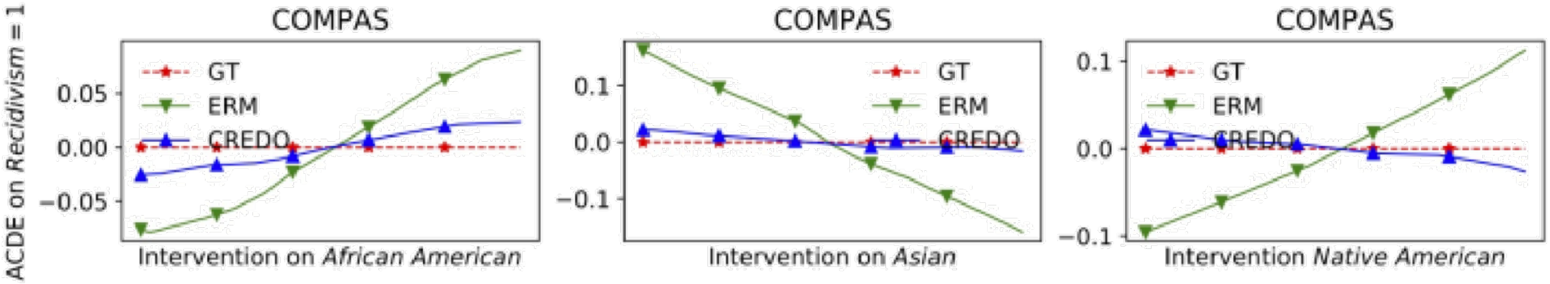
  end

  $j = j + 1$

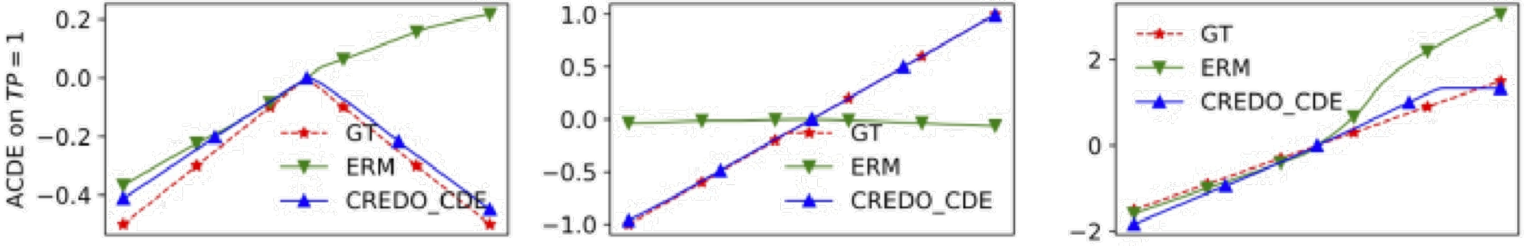
end

return $\frac{1}{N} \sum_{j=1}^{N} max\{0, ||\nabla_j f \odot M - \delta G^j||_1 - \epsilon\}$

Sample Results



MEHRA Dataset



CREDO shows promising performance in matching causal domain priors with no significant impact on model accuracy/training time

Viewing from Different Perspectives: Our Efforts

Post-hoc Explainability

Non-Causal

* [WACV 2018] GradCAM++: Generic method for visual explanations for CNN models
* [IEEE Trans on Biometrics 2021] Canonical saliency maps for face recognition/processing models
* [AISTATS 2022] Submodular ensembles of attribution methods

Causal

* [ICML 2019] Causal attributions in neural networks

Explainability in Deep Learning

Intrinsic Interpretability

Non-Causal

* [CVPR 2022] Ante-hoc explainability via concepts
* [CVPR 2022] Transferring concepts in knowledge distillation tasks

Causal

* [AAAI 2022] Causally disentangled representations
* [CVPR'W 2021] Dataset for causal representation learning
* [WACV 2022] Mitigating bias through causal perspectives
* [arXiv 2022] Causal regularizers

Complementarity of explanations and robustness [AAAI 2021, NeurIPS 2021]

Thank you!

Questions?

vineethnb@cse.iith.ac.in

http://www.iith.ac.in/~vineethnb