Causality in Explainable AI

Motivation and Methods

11 Aug 2023

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

आई आई टी हैदराबाद
IIT Hyderabad

TL; DR

What's this talk about?

Explainability an increasingly core requirement of deployed AI/ML systems;

Actionable and useful explanations are causal ones.

How do the two sit together?

Our Group's Research

Explainable, Robust DL

- *Saliency Maps (Grad-CAM++) and Attributions*, AISTATS 2022, IEEE TBIOM 2021, WACV 2018
- *Causality in NNs*, ICML 2022, AAAI 2022, WACV 2022, ICML 2019
- *Antehoc Interpretability*, CVPR 2022
- *Attributional and Adversarial Robustness*, 2021, ECCV 2020, AAAI 2021

Learning with Limited Labeled Data

- *Continual Learning*, CVPR 2022, WACV 2022, NeurIPS 2020, TPAMI 2021
- *Open-world Learning*, CVPR 2021
- *Few-shot/Zero-shot Learning*, WACV 2021, 2020, CVPR 2019
- *Deep Generative Models*, WACV 2022, CVPR 2018, ICCV 2017

Thesis:

Towards learning robust reliable systems in evolving environments

Deep Learning,
Machine Learning,
Computer Vision

Our Group's Research

Explainable and Robust Learning

- ***Saliency Maps (Grad-CAM++) and Attributions***, AISTATS 2022, IEEE TBIOM 2021, WACV 2018
- ***Causality in NNs***, ICML 2022, AAAI 2022, WACV 2022, ICML 2019, CVPRW 2021
- ***Antehoc Interpretability***, CVPR 2022
- ***Attributional and Adversarial Robustness***, NeurIPS 2021, ECCV 2020, AAAI 2021

Learning in Data/Label-Deficient Environments

- ***Continual Learning***, CVPR 2022, WACV 2022, NeurIPS 2020, TPAMI 2021
- ***Open-world Learning***, CVPR 2021
- ***Few-shot/Zero-shot Learning***, WACV 2021, WACV 2020, CVPR 2019
- ***Deep Generative Models***, WACV 2022, CVPR 2018, ICCV 2017

Deep Learning,
Machine Learning,
Computer Vision

On the Layerwise Hessian of Deep Neural Network Models, **AAAI 2021;** Submodular Batch Selection for Training Deep Neural Networks, **IJCAI 2019;** On Noise and Optimality in Neural Networks, **ICML 2018 Workshops**

Explainability in AI: An Increasing Need

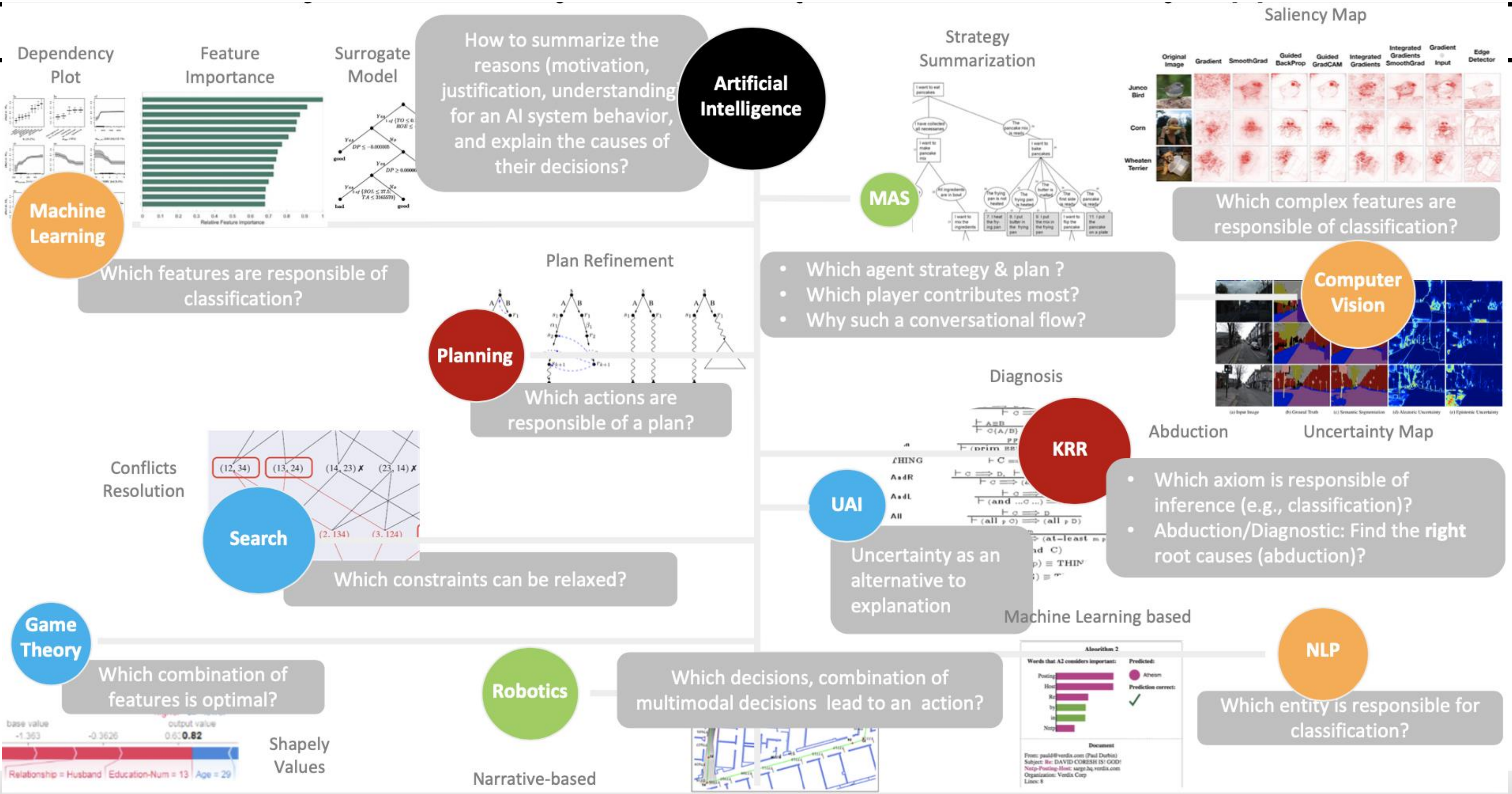
European Union's General Data Protection Regulation (GDPR)

"……a business using personal data for automated processing must be able to explain how the system makes decisions. See Article 15(1)(h) and Recital 71 of GDPR."

Algorithmic Accountability Act 2019: *Requires companies to provide an assessment of risks posed by an automated decision system to privacy or security and the risks that contribute to inaccurate, unfair, biased, or discriminatory decisions impacting consumers*

Right to Explanation:
https://en.wikipedia.org/wiki/Right_to_explanation

Explainable ML: What is being done?



Source: https://xaitutorial2021.github.io

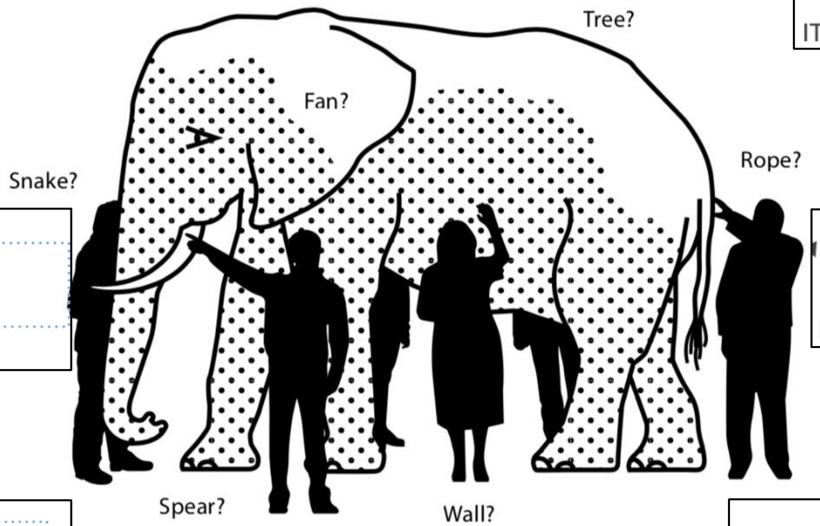
The Elephant in the Room

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

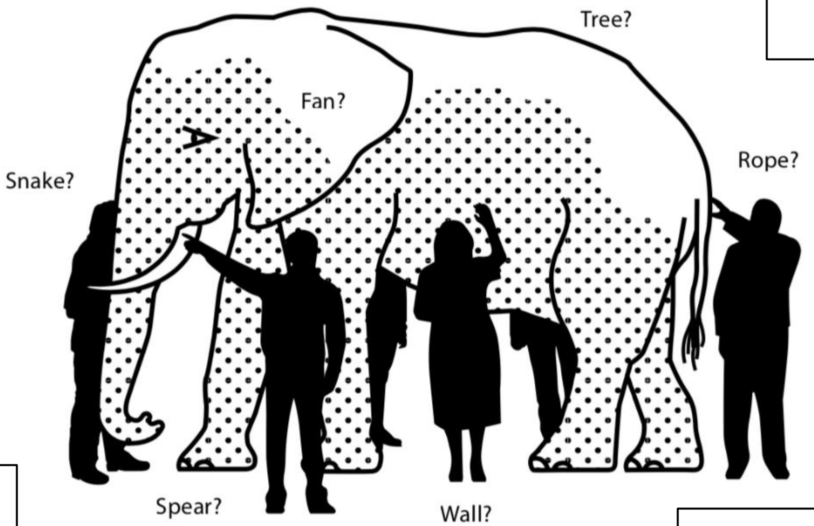
What is it really?

Technical: For those who build it

Post-hoc explainable (vs) Intrinsically interpretable

Transparency (vs) Reasoning

Causal (vs) Correlational associations



Global (vs) Local explanations

Model-agnostic (vs) Model-specific approaches

Attributions (vs) Actionable Explanations

Feature-level (vs) Latent Concept-level Explanations

Viewing XAI 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

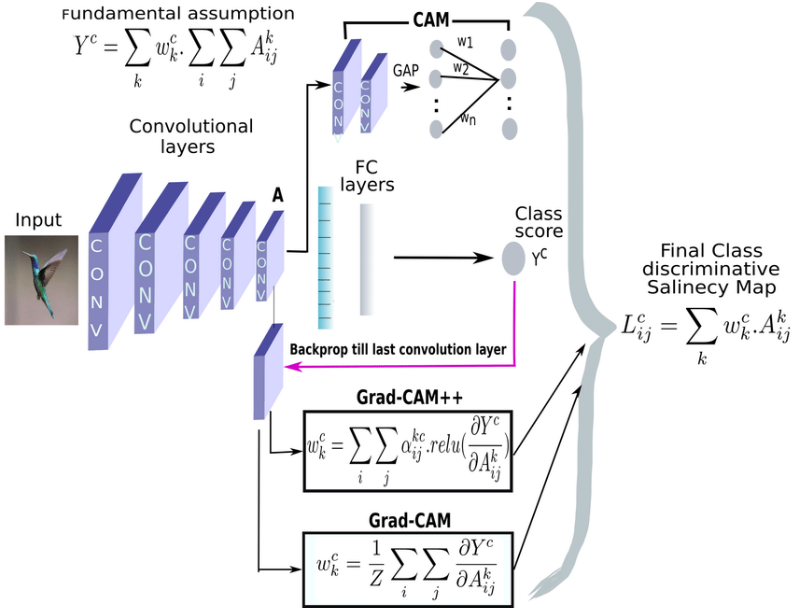
* [AAAI 2022] Causally disentangled representations
* [CVPR'W 2021] Dataset for causal representation learning
* [WACV 2022] Mitigating bias through causal perspectives
* [ICML 2022] Causal regularizers

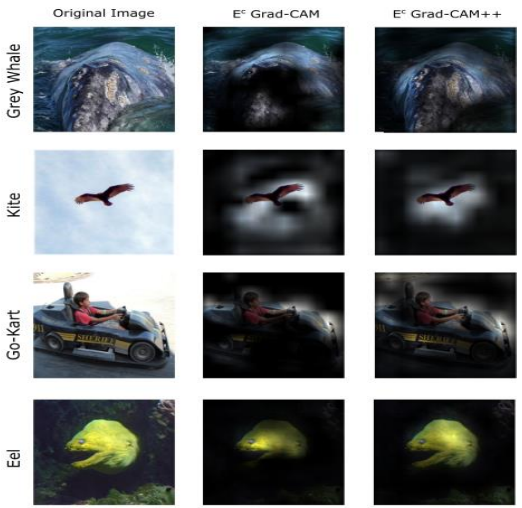
Complementarity of explanations and robustness [AAAI 2021, NeurIPS 2021]

Grad-CAM++

WACV 2018

A pixel-level weighting strategy while computing gradients for explanations





Available on arXiv,
code on Github

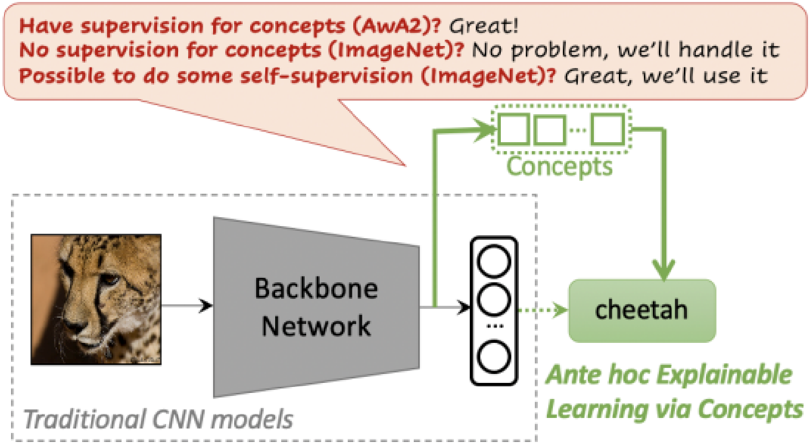
Has been used for:

- Explaining COVID-19 diagnosis in chest X-ray images
- Finding defective cells in solar arrays
- Explaining cancer prediction on gene expression data
- Identification of pathogens in tomograms
- Leaf counting, Genus classification in plant images
- …

~1500 citations at this time

Ante-hoc Explainability via Concepts

CVPR 2022



Dataset	Baselines SENN[4]	CBM[5]	OURS w/o sup	w sup
CIFAR10	84.50	NA	**91.68**	NA
ImageNet	58.55	NA	**65.09**	NA
AwA2	76.41	81.61	81.04	**85.70**
CUB-200	58.81	64.17	63.05	**65.28**

Accuracy (in %) using ResNet18 architecture as concept (or base) encoder

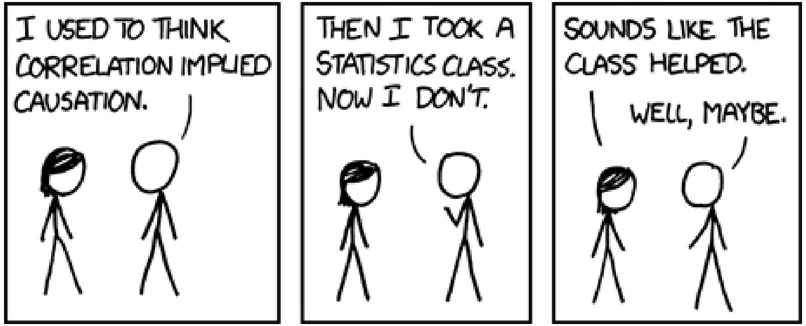
- Learn latent concept-based explanations implicitly during training
- Append explanation generation module on any basic network and jointly train whole module.
- Provides explanations that are global (concepts that are most activated on a dataset or a class) or local (concepts that are most activated for prediction on given input image).
- Can be easily integrated with existing backbone networks.
- Works with different levels of supervision

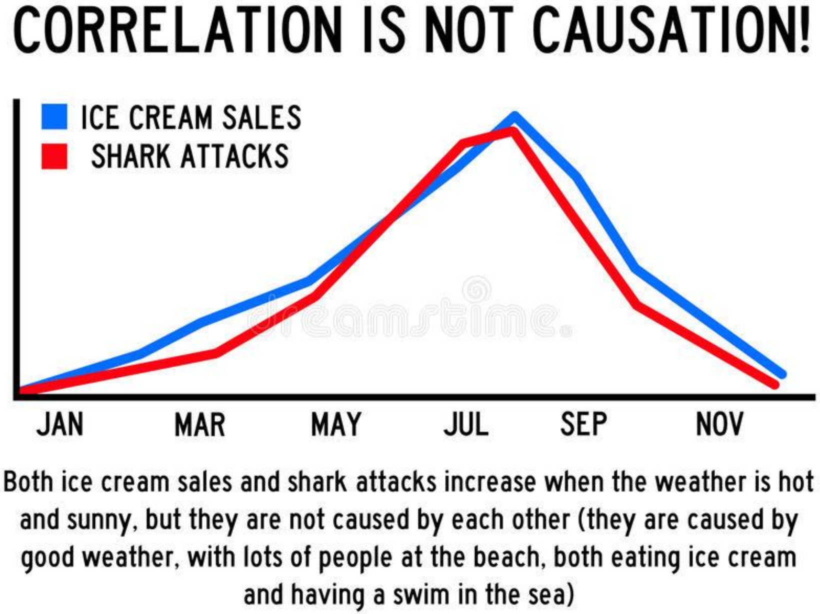
Causal XAI: What and Why?

Alright, alright – but why causal?
What are causal explanations?

Causation vs Correlation

Is feature correlation of input to output a true explanation?





Simpson's Paradox

- Consider vaccines for COVID-19
  - **Treatment T (Vaccine):** A (0) or B (1)
  - **Condition C:** Mild (0) or Severe (1)
  - **Outcome Y:** Alive (0) or Dead (1)

	Mild	Severe	Total
A	15%(210/1400)	30%(30/100)	16%(240/1500)
B	10%(5/50)	20%(100/500)	19%(105/550)
	$\mathbb{E}[Y \mid T, C=0]$	$\mathbb{E}[Y \mid T, C=1]$	$\mathbb{E}[Y \mid T]$

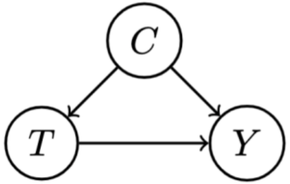
Mortality Rate Table

	Total
A	16%(240/1500)
B	19%(105/550)
	$\mathbb{E}[Y \mid T]$

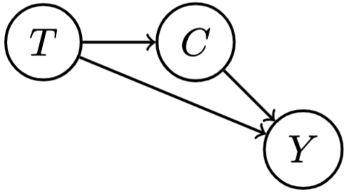
Which treatment to choose?

Now, which treatment to choose?

Depends on the causal graph!



Treatment B

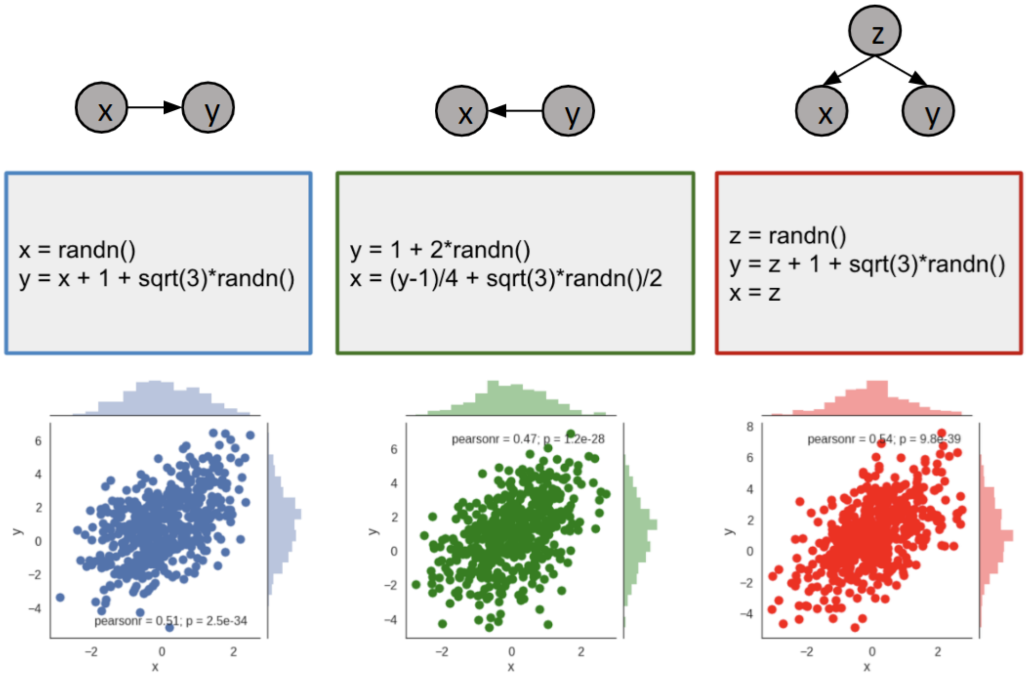


Treatment A

Let's see a different perspective

Core objective in supervised ML tasks – how are **x** (data) and **y** (labels) related?

Same joint distribution **p(x,y)** can be generated by different variable relations!



Credit: Gautam Gare, CMU

Let's see a different perspective

Core obje… …s

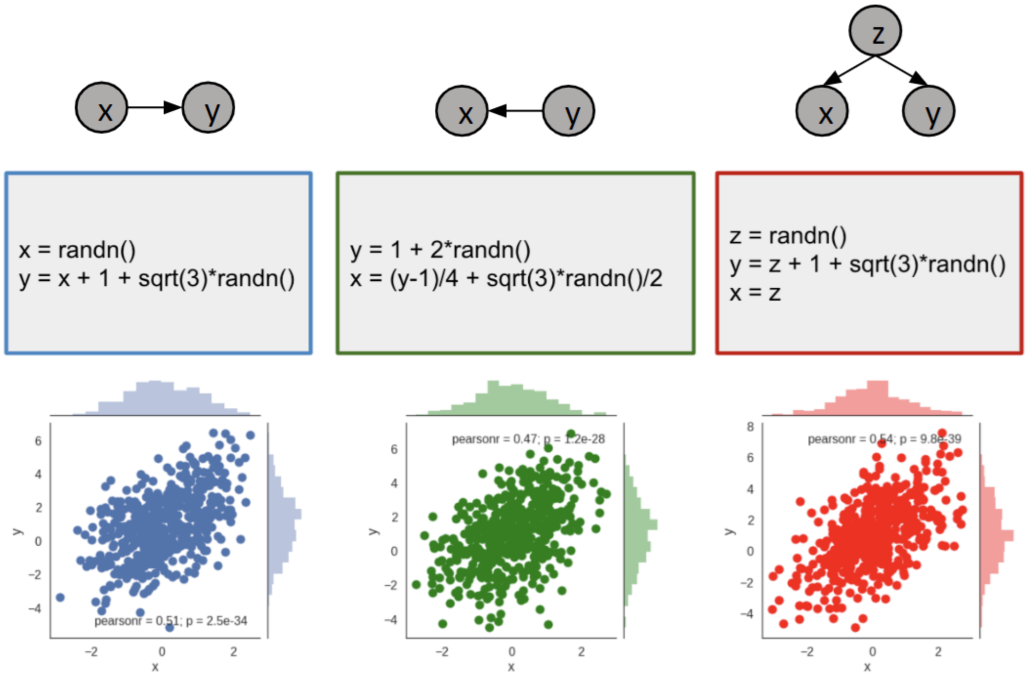
– how…

Why does it matter, so long it fits the data perfectly?

Same jo…

be gene…

It only fits this data perfectly

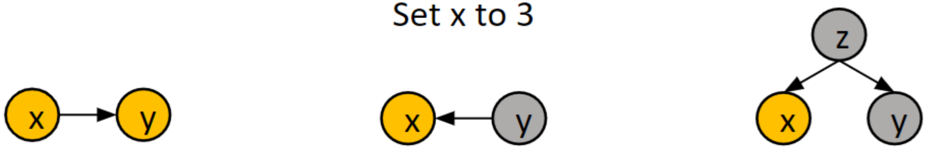


Credit: Gautam Gare, CMU

Evaluate on data from other distributions?

Let's change a variable

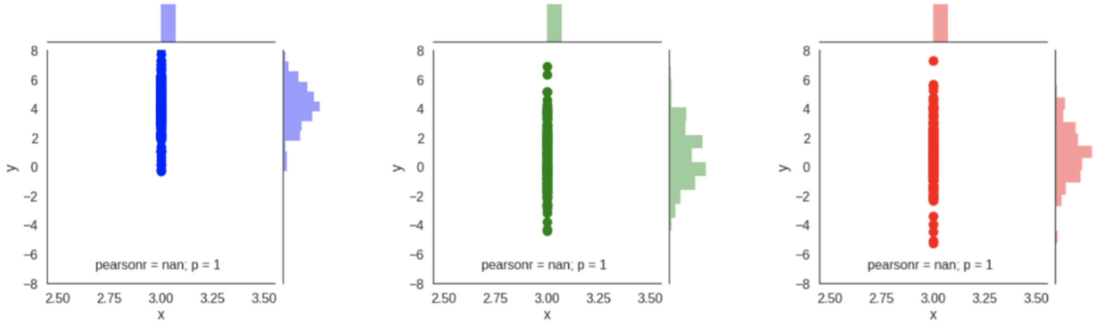
Set x to 3




```
x = randn()
x = 3
y = x + 1 + sqrt(3)*randn()
x = 3
```

```
y = 1 + 2*randn()
x = 3
x = (y-1)/4 + sqrt(3)*randn()/2
x = 3
```

```
z = randn()
x = 3
x = z
x = 3
y = z + 1 + sqrt(3)*randn()
x = 3
```

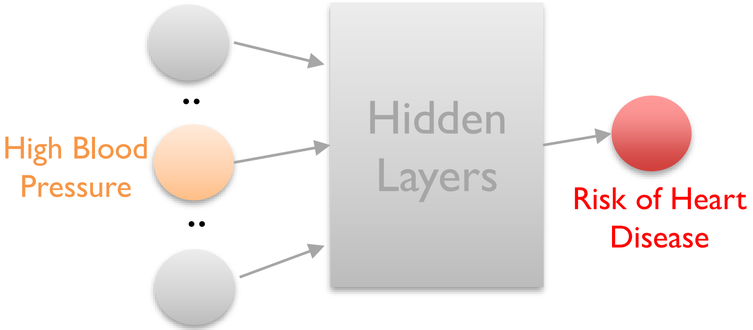


Knowing the true causal relationships makes a difference!

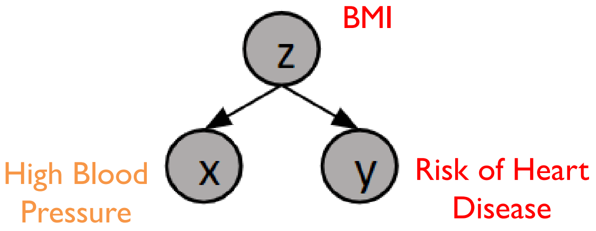
Credit: Gautam Gare, CMU

Does this matter in XAI?

- Training a NN model to predict risk of heart disease
- Post-hoc explanations focus on data correlations NN has learned to provide input-output attributions



- …but what if the causal graph had a "confounder"? Would the explanation address the problem?



He et al, Causal effects of cardiovascular risk factors on onset of major age-related diseases: A time-to-event Mendelian randomization study, Exp Gerontol. 2018

More generally..

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.

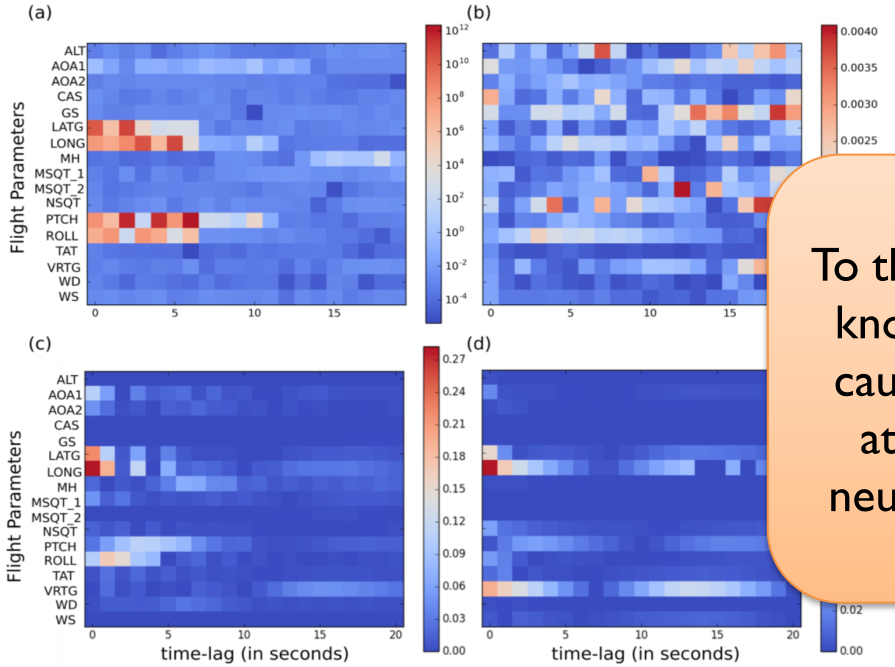
Judea Pearl, The Seven Tools of Causal Inference with Reflections on Machine Learning, 2018
Judea Pearl, The Book of Why: The New Science of Cause and Effect, 2018

Causal XAI: How?

How can we get causal explanations?
How can we integrate causal perspectives into model explanations?

Causal Attributions in Neural Networks

ICML 2019



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

Joint work with:

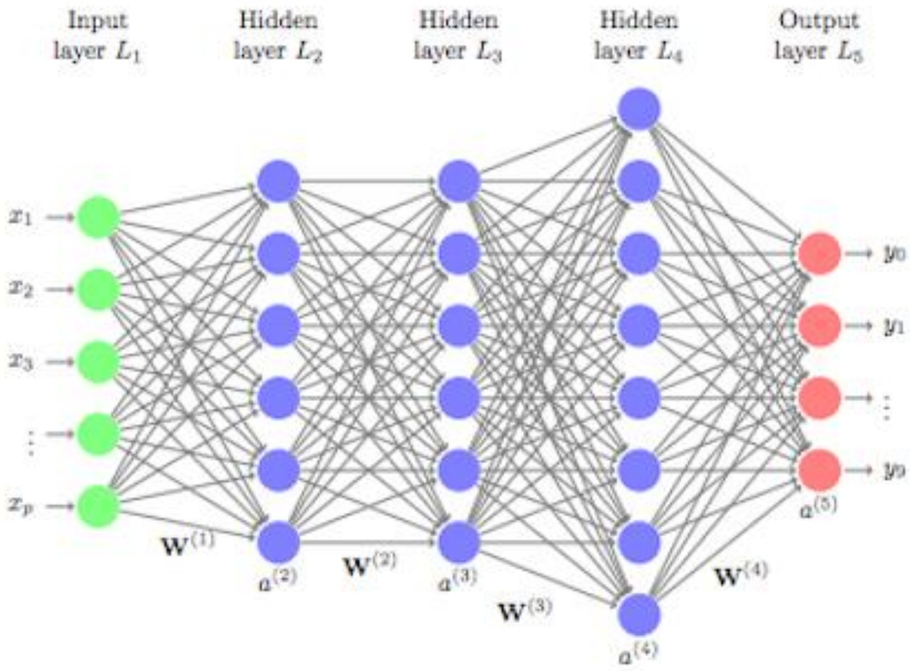
Aditya Chattopadhyay

Piyushi Manupriya

Anirban Sarkar

आई आई टी हैदराबाद
IIT Hyderabad

Do NNs Learn Causal Relationships?



Consider a trained NN model. Did it learn causal relationships between input and output?

SCMs and Causal Effect

Preliminaries

Structural Causal Model

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

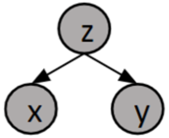
Endogenous variables

Exogenous variables

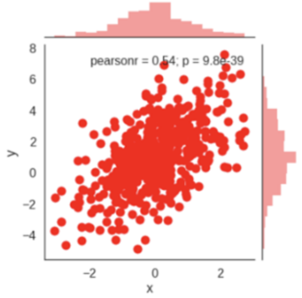
Causal Functions

Distribution of U

Example:

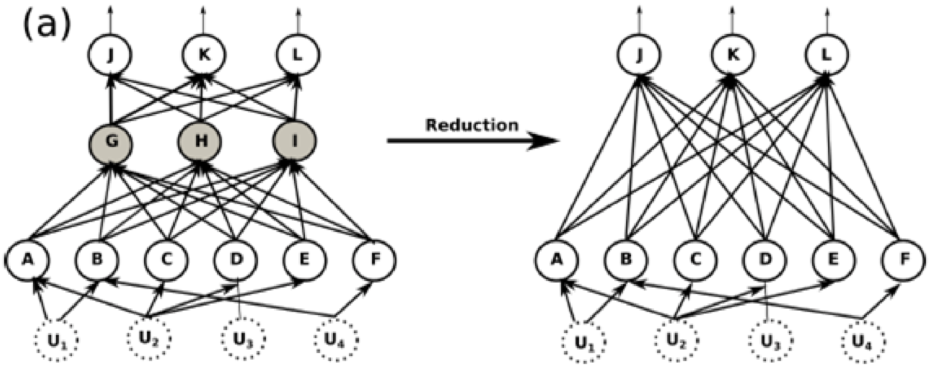



```
z = randn()
y = z + 1 + sqrt(3)*randn()
x = z
```



Neural Network as an 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 l
- f_i – corresponding causal functions

Defining Causal Effect

How to compute causal effect for a trained NN?

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

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

- **Connection to Attribution:** Effect of an input feature on prediction function's output
- Existing attribution/explanation 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

Computing Average Causal Effect in NN

General case (continuous variables):

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

Interventional expectation:
How to compute?

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

How to define and compute?

- Can come from domain knowledge
- Else, we use $\mathbb{E}_{x_i}[\mathbb{E}_y[y|do(x_i=\alpha)]]$

  the average ACE across all x_i

Computing ACE

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

Let: $y = f'_y(x_1, x_2, ..., x_k)$ $\qquad \mu_j = \mathbb{E}[x_j|do(x_i = \alpha)]\forall x_j \in l_1$

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

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

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

Proposition 2. Given an l-layer feedforward neural network $N(l_1, l_2, ...l_n)$ with l_i denoting the set of neurons in layer i and its corresponding reduced SCM $M'([l_1, l_n], U, f', P_U)$, the intervened input neuron is d-separated from all other input neurons.

$\rightarrow$

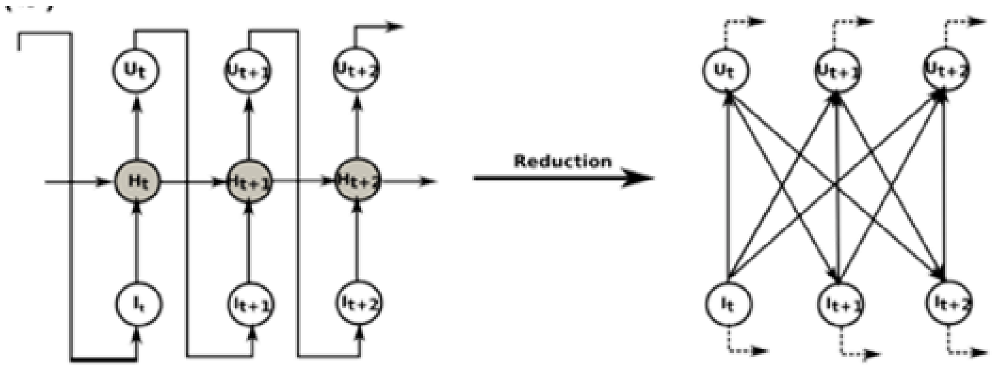
- Given an intervention on a particular variable, the probability distribution of all other input neurons doesn't change, i.e. for $x_j \neq x_i$

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

- Interventional means and covariances of non-intervened neurons same as observational means and covariances; can be pre-computed!

Only for feedforward NNs?

Recurrent neural network

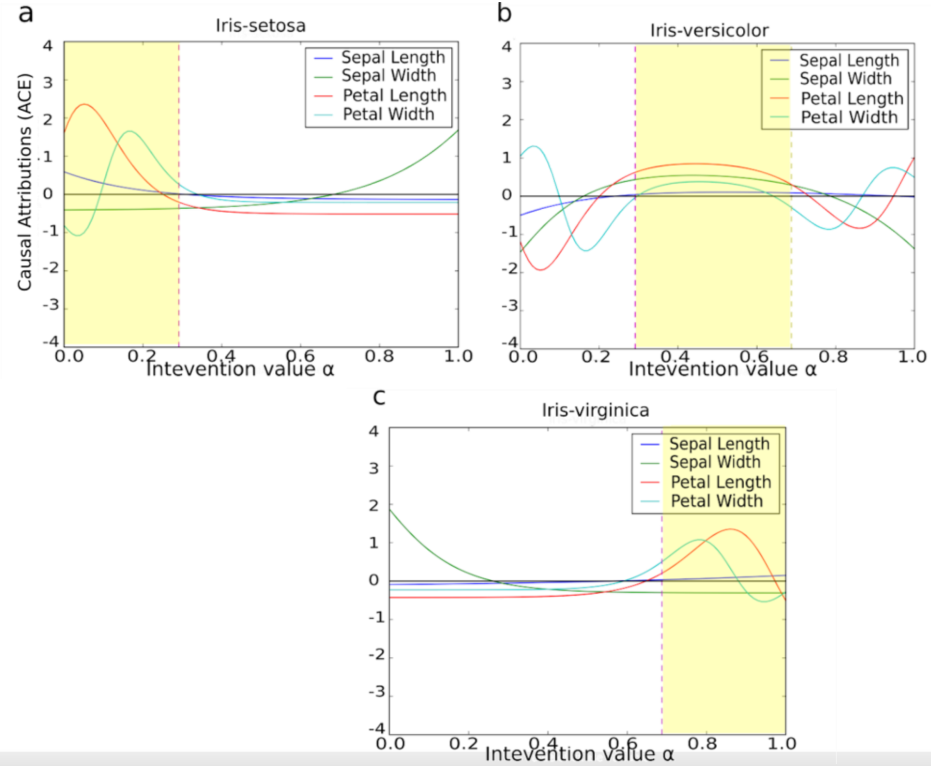


Depends on a particular RNN architecture.

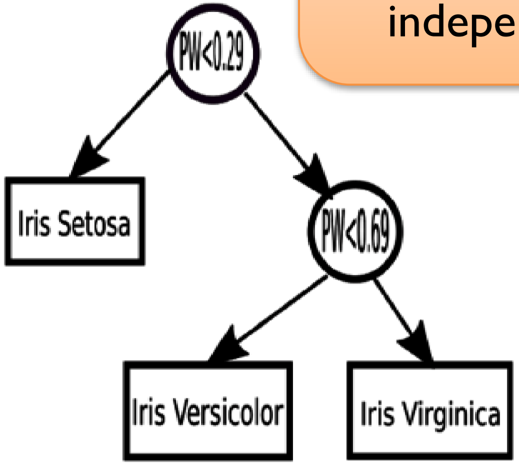
Where output does not feed into input, same idea can be used

Results

Iris Dataset



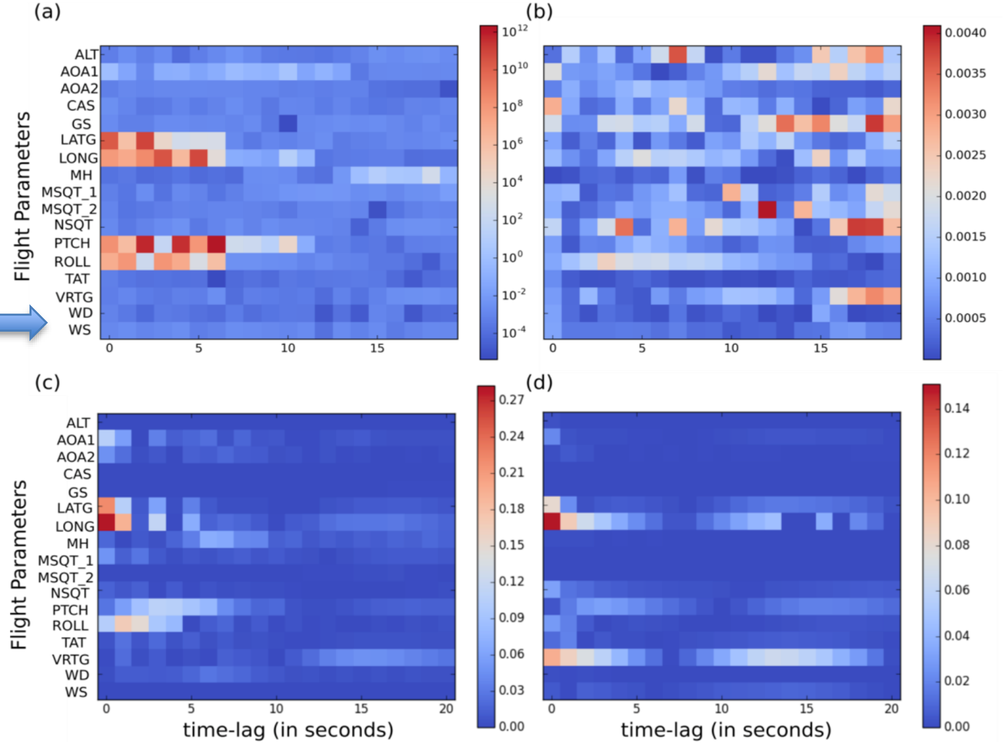
ACE values match a decision tree learned independently



Results

Aircraft Data (NASA Dashlink Dataset)

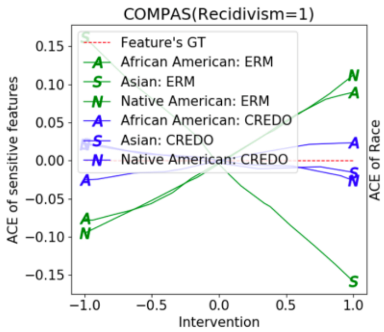
FDR report:"….*due to slippery runway, the pilot could not apply timely brakes, resulting in a steep acceleration in the airplane post-touchdown…*"

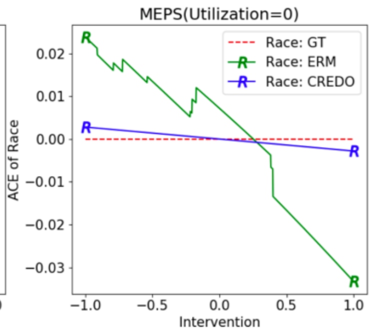


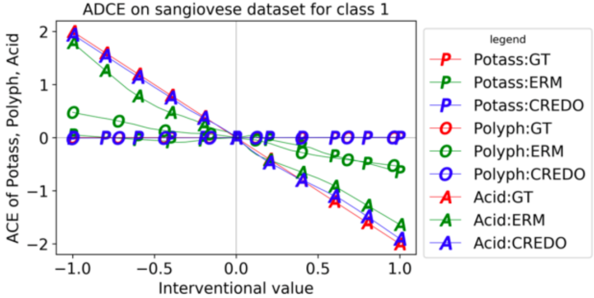
arXiv:
https://arxiv.org/abs/1902.02302
Code:
https://github.com/Piyushi-0/ACE

Causal Regularization with Domain Priors

ICML 2022



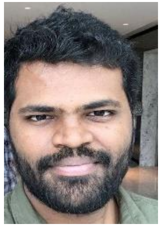




To the best of our knowledge, first effort to integrate causal knowledge for attribution in neural networks

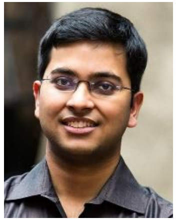
Joint work with:

Gowtham Reddy A



Sai Srinivas K

Amit Sharma



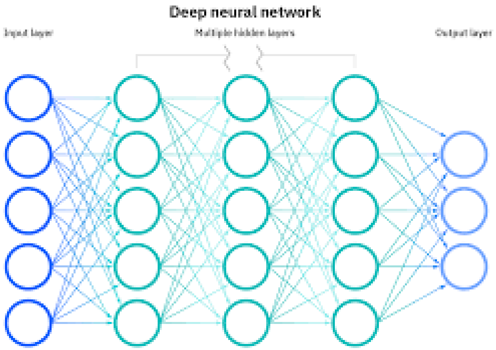
Microsoft Research

आई आई टी हैदराबाद
IIT Hyderabad

Do NNs Learn Causal Relationships?

ICML 2019 and ICML 2022

Consider a trained NN model. Did it learn causal relationships between input and output?



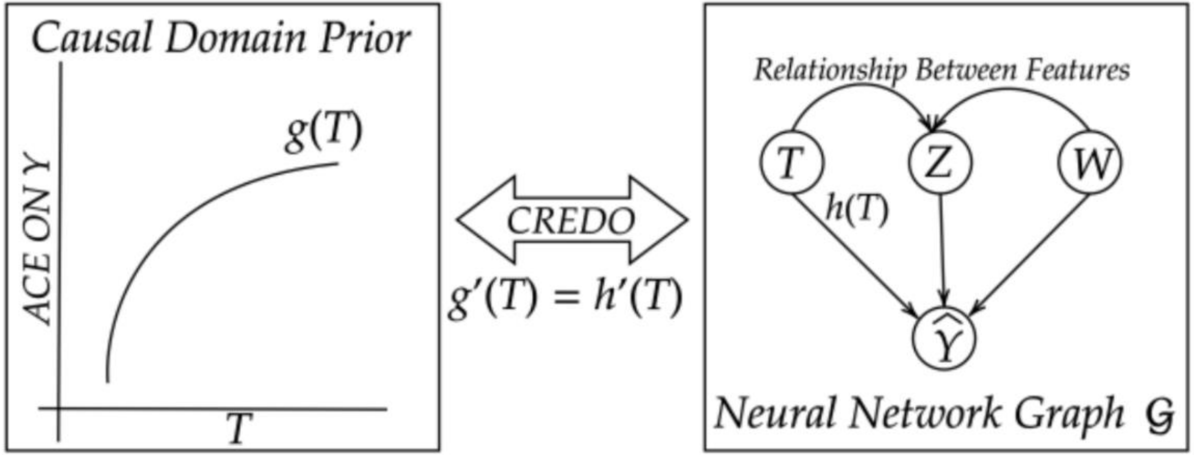
If we had access to prior causal relationships, can we integrate them while training NN models?

Causal Attributions in Neural Networks
ICML 2019

Causal Regularization with Domain Priors
ICML 2022

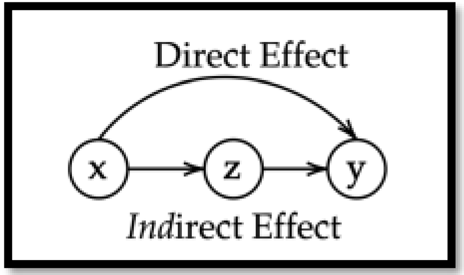
Key Idea

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



We handle three kinds of causal effect NN models in this work:

- Controlled direct effect
- Natural direct effect
- Total causal 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: $NN-ACDE^{\hat{Y}}_{t,PA^{\hat{Y}}=\alpha} := \mathbb{E}_U[\hat{Y}_{t,PA^{\hat{Y}}=\alpha}] - \mathbb{E}_U[\hat{Y}_{t^*,PA^{\hat{Y}}=\alpha}] = \hat{Y}_{t,PA^{\hat{Y}}=\alpha} - \hat{Y}_{t^*,PA^{\hat{Y}}=\alpha}$.

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

Pearl, Causality: Models, Reasoning and Inference, 2003

Identifiability in Causality

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 $ACDE_t^{\hat{Y}} = \mathbb{E}_{PA^{\hat{Y}}}\left[\hat{Y}|t, PA^{\hat{Y}}\right] - \mathbb{E}_{PA^{\hat{Y}}}\left[\hat{Y}|t^*, PA^{\hat{Y}}\right]$.

Identifiability:
the condition that permit to measure causal quantity from observed data

$$
\begin{aligned}
ACDE_t^{\hat{Y}} &= \mathbb{E}_{Z,W,U}[\hat{Y}_{t,Z,W}] - \mathbb{E}_{Z,W,U}[\hat{Y}_{t^*,Z,W}]] \\
&= \mathbb{E}_{Z,W}[\hat{Y}_{t,Z,W}] - \mathbb{E}_{Z,W}[\hat{Y}_{t^*,Z,W}] \\
&= \mathbb{E}_{Z,W}[\hat{Y}|t, Z, W] - \mathbb{E}_{Z,W}[\hat{Y}|t^*, Z, W]
\end{aligned}
$$

Regularizing for Causal Effect

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: $\frac{\partial^n ACDE_t^{\hat{Y}}}{\partial t^n} = \mathbb{E}_{PA^{\hat{Y}}}\left[\frac{\partial^n [\hat{Y}(t, PA^{\hat{Y}})]}{\partial t^n}\right]$.

$$\frac{\partial^n ACDE_t^{\hat{Y}}}{\partial t^n} = \frac{\partial^n [\mathbb{E}_{Z,W}[\hat{Y}|t, Z, W] - \mathbb{E}_{Z,W}[\hat{Y}|t^*, Z, W]]}{\partial t^n}$$

$$= \frac{\partial^n [\mathbb{E}_{Z,W}[\hat{Y}|t, Z, W]]}{\partial t^n} (\because t^* \; is \; a \; constant)$$

$$= \mathbb{E}_{Z,W}\left[\frac{\partial^n [\hat{Y}(t, Z, W)]}{\partial t^n}\right]$$

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

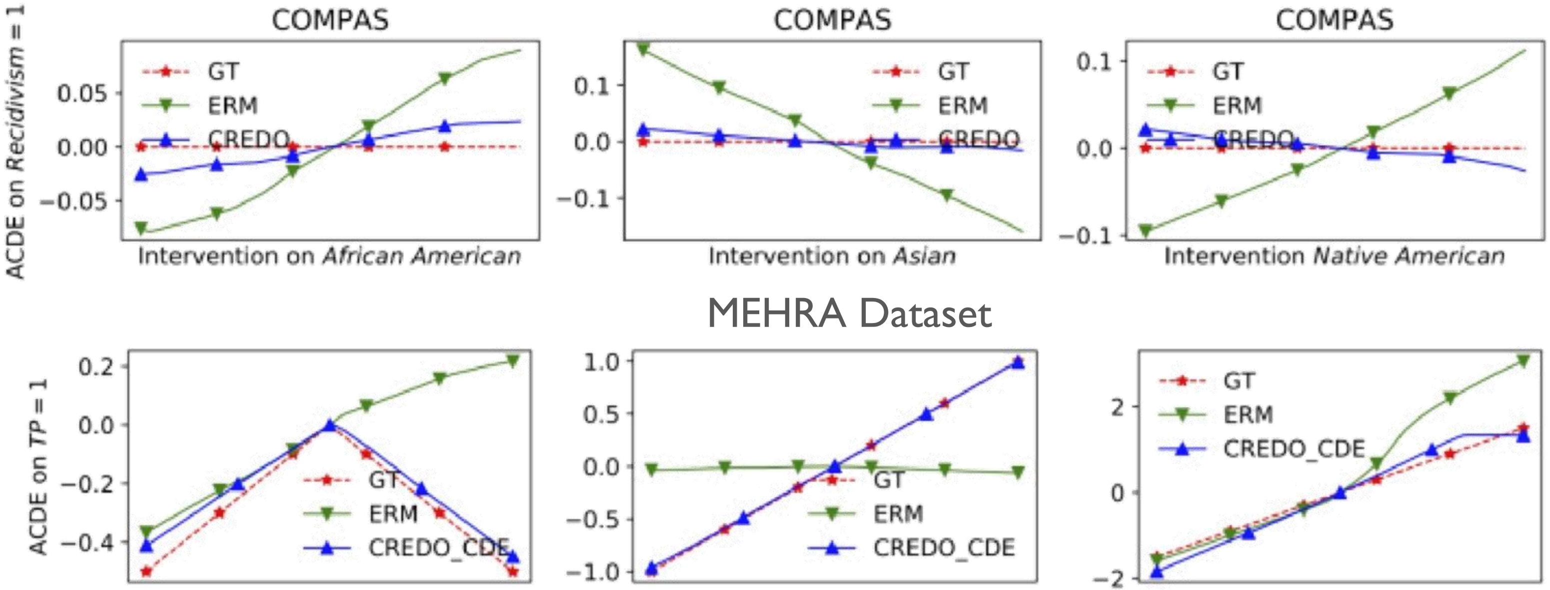
Initialize: $j = 1, \delta G^j = \mathbb{0}_{c \times d} \forall j = 1, \ldots, N$, $M = \mathbb{0}_{c \times d}$

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



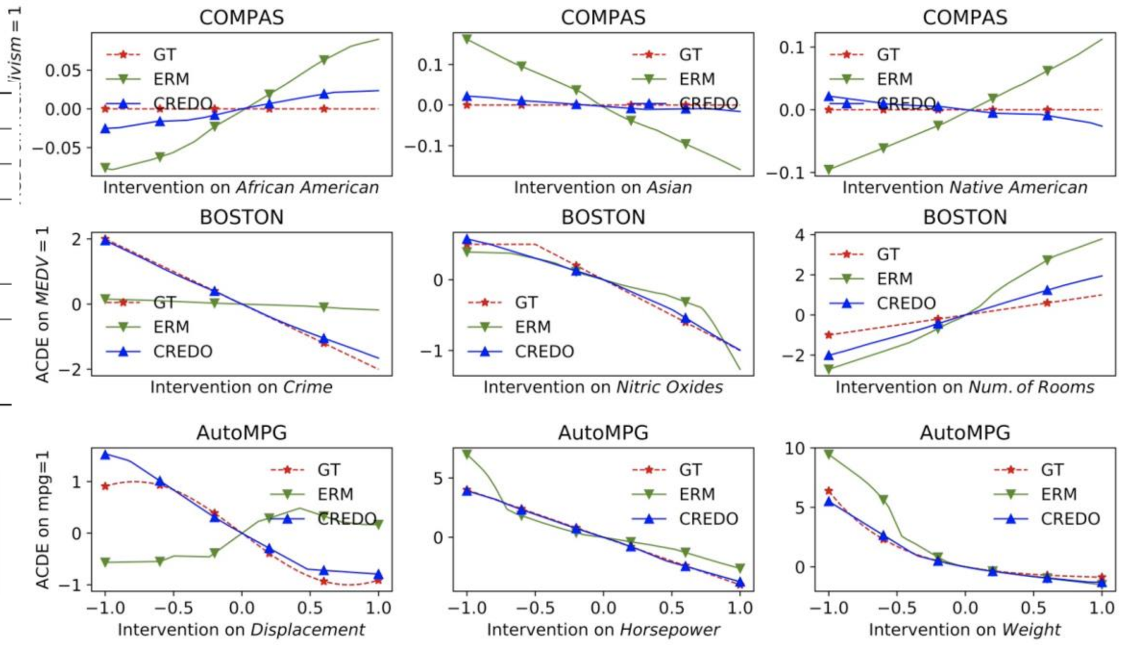
CREDO shows promising performance in matching causal domain priors with no significant impact on model accuracy/training time

Sample Results

Causal graph unknown

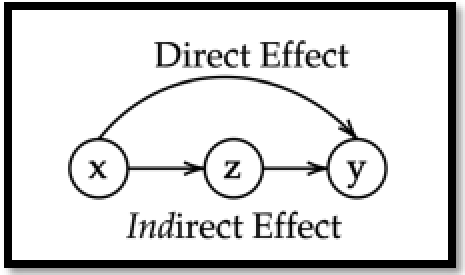
Feature	RMSE		Frechet Score		Corr. Coeff.	
	ERM	CREDO	ERM	CREDO	ERM	CREDO
COMPAS ($\lambda_1 = 5$) (ERM test accuracy is 67.90%, CREDO test accuracy is 67.09%)						
African American	0.055	**0.016**	0.088	**0.025**	-	-
Asian	0.092	**0.018**	0.162	**0.021**	-	-
Native American	0.059	**0.011**	0.109	**0.025**	-	-
AutoMPG ($\lambda_1 = 1.5$) (ERM test accuracy is 88.6%, CREDO test accuracy is 87.34%)						
Displacement	1.144	**0.212**	**0.566**	1.524	-0.945	**0.977**
Horsepower	1.036	**0.081**	6.978	**3.908**	0.922	**0.999**
Weight	1.780	**0.25**	9.453	**5.510**	0.986	**0.992**

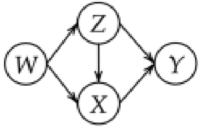
arXiv:
https://arxiv.org/abs/2111.12490



Causal Attributions: Going Beyond Direct Effects

arXiv Preprint 2303.13850





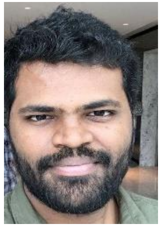
$$W \leftarrow Uniform(0,1)$$
$$Z \leftarrow 2W + \mathcal{N}(0, 0.1)$$
$$X \leftarrow 2W - Z + \mathcal{N}(0, 0.1)$$
$$Y \leftarrow 3X + e^{3Z} + \mathcal{N}(0, 0.1)$$

Table 1: Synthetic Data 1

	Feature	**IG** [S 2017]	**CA** [AC 2019]	**CREDO** [SK 2022]	**Ours**
		Synthetic Data 1			
RMSE (↓)	W	0.869	0.869	0.835	1.114
	Z	0.569	0.569	0.804	0.373
	X	0.000	0.000	0.229	0.314
	Average	0.479	**0.326**	0.622	0.618
Frechet (↓)	W	1.000	1.000	1.000	1.000
	Z	1.000	1.000	1.883	0.883
	X	0.000	0.000	0.397	0.352
	Average	**0.667**	**0.667**	1.109	0.745

Joint work with:

Gowtham Reddy A



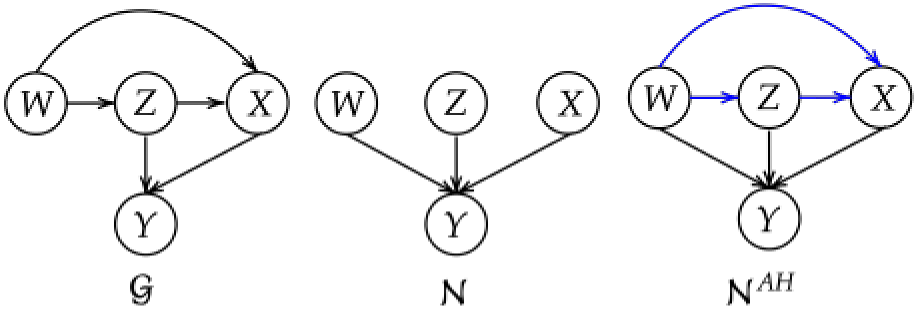
Saketh Bachu

Varshaneya

Honeywell

आई आई टी हैदराबाद
IIT Hyderabad

Going Beyond Direct Effects: Key Idea



We introduce connections among input features to capture underlying causal relationships to learn indirect causal attributions of inputs on y

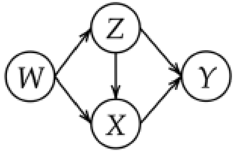
Identifiability and Training Algorithm

Proposition 4.1: *Given a neural network $\mathcal{N}$ with directed edges among input features $x_1, \ldots, x_n$ denoting causal relationships among the features in the underlying causal graph $\mathcal{G}$, the $AICE^{y}_{x_i}$ of an input feature x_i on an output neuron y is identifiable in $\mathcal{N}$.*

Algorithm 1 Training Algorithm for Proposed $\mathcal{N}^{AH}$

Input: Causal graph $\mathcal{G}$, $\mathcal{D} = \{(x^i_1, \ldots, x^i_n, y^i)\}^m_{i=1}$, $l_0 =$ edges among $\{x_1, \ldots, x_n\}$
Output: Trained $\mathcal{N}^{AH}$
for each epoch **do**
  for phase in [freeze, full] **do**
    if phase = freeze **then**
      Freeze l_0, train $l_1, \ldots, l_n$ of $\mathcal{N}^{AH}$ using $\mathcal{D}$
    else
      $X^r = \{x_i : pa(x_i) = \emptyset\}$
      Sample $X \setminus X^r$ using l_0, X^r
      Train $l_0, \ldots, l_n$ of $\mathcal{N}^{AH}$ using (X, y).
    end if
  end for
end for
return trained $\mathcal{N}^{AH}$

Results

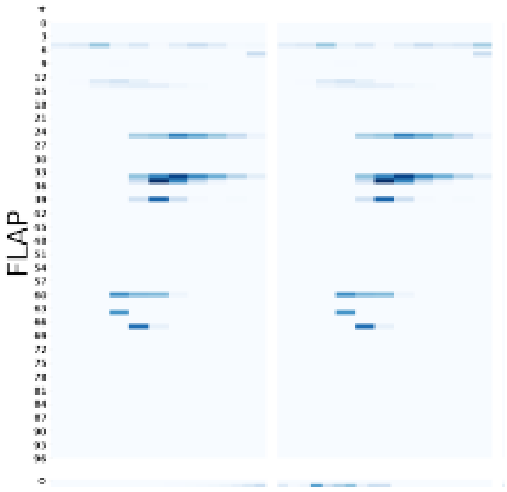


$$W \leftarrow Uniform(0,1)$$
$$Z \leftarrow 2W + \mathcal{N}(0, 0.1)$$
$$X \leftarrow 2W - Z + \mathcal{N}(0, 0.1)$$
$$Y \leftarrow 3X + e^{3Z} + \mathcal{N}(0, 0.1)$$

Table 1: Synthetic Data 1

	Feature	IG [S 2017]	CA [AC 2019]	CREDO [SK 2022]	Ours
		Synthetic Data 1			
RMSE (↓)	W	0.869	0.869	0.835	1.114
	Z	0.569	0.569	0.804	0.373
	X	0.000	0.000	0.229	0.314
	Average	0.479	**0.326**	0.622	0.618
Frechet (↓)	W	1.000	1.000	1.000	1.000
	Z	1.000	1.000	1.883	0.883
	X	0.000	0.000	0.397	0.352
	Average	**0.667**	**0.667**	1.109	0.745

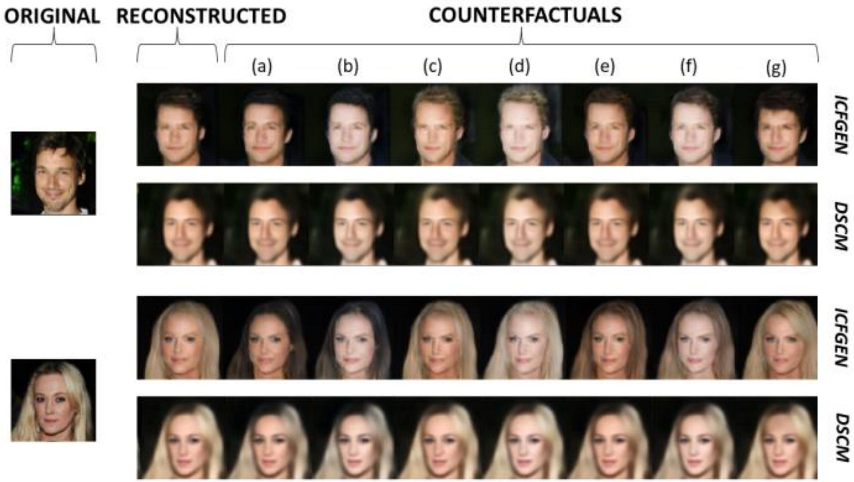
Flight anomaly datasets



arXiv:
https://arxiv.org/abs/2303.13850

Evaluating and Mitigating Bias in Image Classifiers: A Causal Perspective Using Counterfactuals

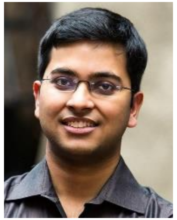
WACV 2022



Joint work with:

Saloni Dash

Amit Sharma



Microsoft Research

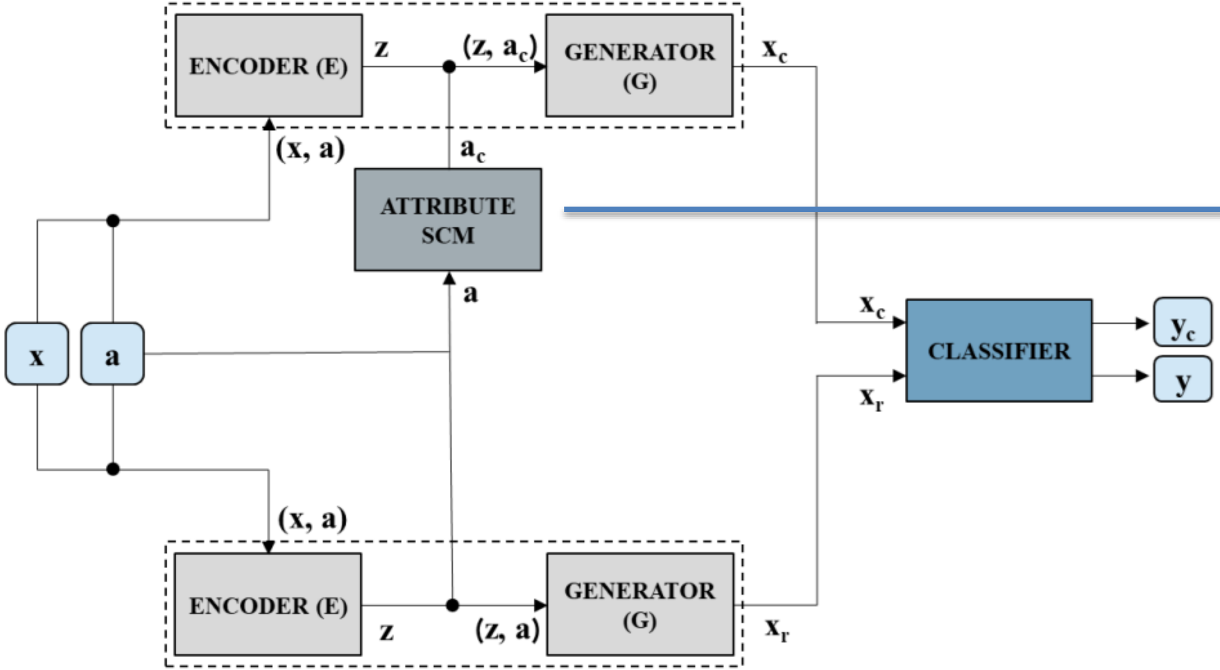
आई आई टी हैदराबाद
IIT Hyderabad

Causal Perspective to Counterfactual Generation

- Existing perspectives to counterfactuals in DL very weak and scattered – not truly causal
- How to integrate a causal perspective in counterfactual generation, and what could be its applications?

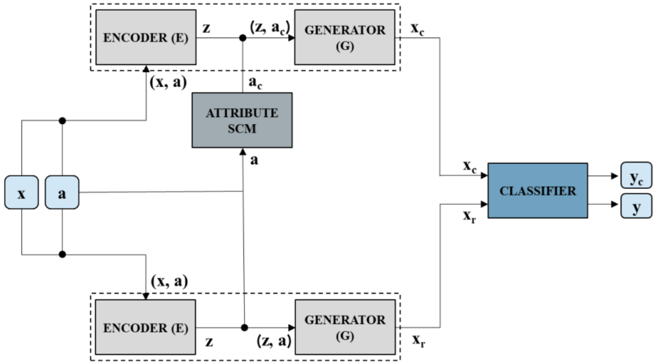
- Image we want to generate the counterfactual for: $\mathbf{x} \in \mathcal{X}$
- Corresponding attributes: $\mathbf{a} = \{a_i\}_{i=1}^{n} \in \mathcal{A}$
  - E.g. Smiling, Brown hair for Celeb-A; Thickness, Intensity for MNIST
- Given $(\mathbf{x}, \mathbf{a})$ goal is to generate a counterfactual image with the attributes changed to $\mathbf{a}_c$

Counterfactual Generation



- Individual local SCMs learned over attributes
- Overall model learned similar to a GAN (using ALI)

Counterfactual Generation



- An encoder $E : (\mathcal{X}, \mathcal{A}) \rightarrow \mathbf{Z}$ infers the latent vector $\mathbf{z}$ from $\mathbf{x}$ and $\mathbf{a}$, i.e. $\mathbf{z} = E(\mathbf{x}, \mathbf{a})$ where $\mathbf{Z} = \mathbf{z} \in \mathbb{R}^m$.
- The Attribute-SCM intervenes on the desired subset of attributes that are changed from $\mathbf{a}$ to $\mathbf{a}'$, resulting in output $\mathbf{a}_c$.
- Generator $G : (\mathbf{Z}, \mathcal{A}) \rightarrow \mathcal{X}$ takes as input $(\mathbf{z}, \mathbf{a}_c)$ and generates a counterfactual $\mathbf{x}_c$, where $\mathbf{z} \in \mathbf{Z} \subseteq \mathbb{R}^m$.

- Abduction
- Action
- Prediction

Applications?

- Evaluating fairness of a classifier

$$\text{bias} = p(y_r \neq y_c)(p(y_r = 0, y_c = 1 | y_r \neq y_c) \\ -p(y_r = 1, y_c = 0 | y_r \neq y_c))$$

$$\Longrightarrow \quad \text{bias} = p(y_r = 0, y_c = 1) - p(y_r = 1, y_c = 0)$$

- Explaining a classifier (in terms of attributes)

$$\mathbb{E}_Y[Y_{\mathbf{a}_i \leftarrow a'} | \mathbf{x}, \mathbf{a}] - \mathbb{E}_Y[Y_{\mathbf{a}_i \leftarrow a} | \mathbf{x}, \mathbf{a}] \\ = y_{\mathbf{a}_i \leftarrow a'} | \mathbf{x}, \mathbf{a} - y_{\mathbf{a}_i \leftarrow a} | \mathbf{x}, \mathbf{a}$$

- Bias mitigation:

Train using $\text{BCE}(y_{true}, \hat{f}(\mathbf{x})) + \lambda \text{MSE}(\text{logits}(\mathbf{x}_r), \text{logits}(\mathbf{x}_c))$

Counterfactual Generation

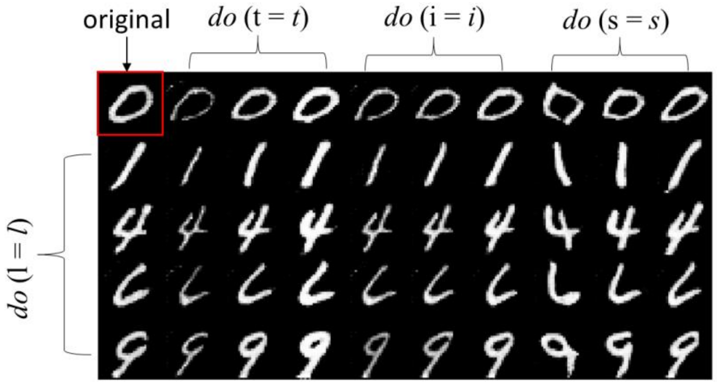


Figure 3: **Morpho-MNIST Counterfactuals.** Top-left cell shows a real image sampled from the test set. Vertically, rows correspond to interventions on the label, $do(l = 1, 4, 6, 9)$. Moving horizontally, columns correspond to interventions on thickness: do (t=1, 3, 5), intensity: do (i = 68, 120, 224), and slant: do (s = -0.7, 0, 1) respectively.

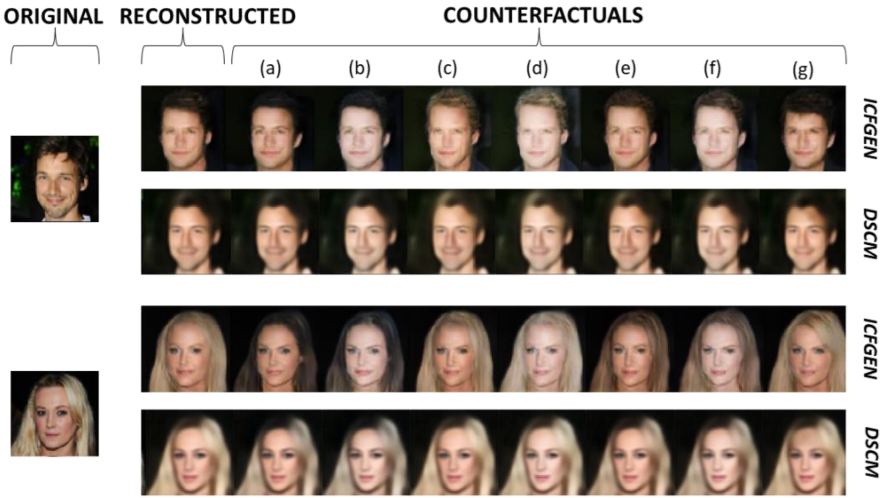


Figure 5: ***ImageCFGen* and DeepSCM Counterfactuals.** (a) denotes do (black hair = 1) and (b) denotes do (black hair = 1, pale =1). Similarly (c) denotes do (blond hair = 1); (d) denotes do (blond hair = 1, pale = 1); (e) denotes do (brown hair = 1); (hf denotes do (brown hair = 1, pale = 1); and (g) denotes do (bangs = 1).

Counterfactual Generation

	p($a_r \neq a_c$)	p($0 \rightarrow 1$)	bias
horizontal_flip	0.073	0.436	-0.009
brightness	0.192	0.498	-0.001
black_h	0.103	0.586	0.018
black_h, pale	0.180	0.937	**0.158**
blond_h	0.115	0.413	- 0.02
blond_h, pale	0.155	0.738	**0.073**
brown_h	0.099	0.704	0.041
brown_h, pale	0.186	0.942	**0.164**
bangs	0.106	0.526	0.005

Table 3: **Bias Estimation.** Bias values above a threshold of 5% are considered significant.

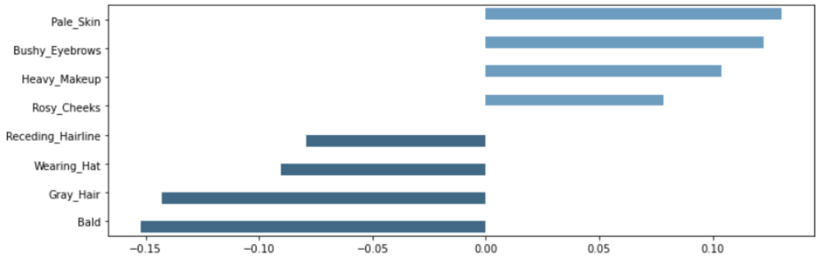


Figure 7: **Explaining a Classifier.** Attribute ranking of top 4 positive and top 4 negative influential attributes.

Classifying a face as attractive

Bias Mitigation. Using generated CFs reduces bias to 0.032 for black hair and pale, and 0.012 for brown hair and pale

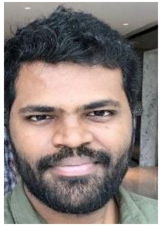
For more details:
https://arxiv.org/abs/2009.08270

On Causally Disentangled Representations

AAAI 2022

Joint work with:

Gowtham Reddy A



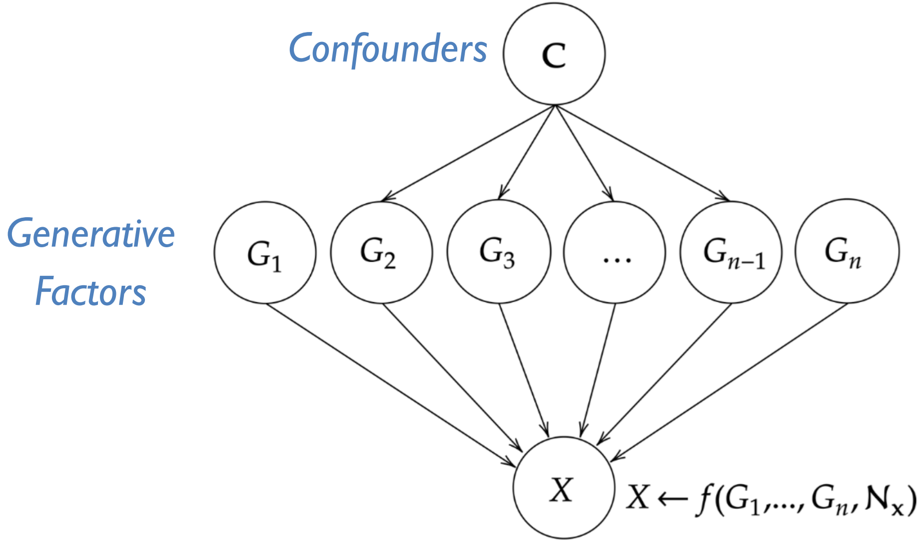
Benin Godfrey

आई आई टी हैदराबाद
IIT Hyderabad

Causal Disentanglement

Our Work

- Disentanglement has been a topic of recent interest – however most existing methods assume independence among latent variables (generative factors)
- We present two evaluation metrics based on the properties of causally disentangled LVMs
- We develop a new weakly supervised disentanglement algorithm



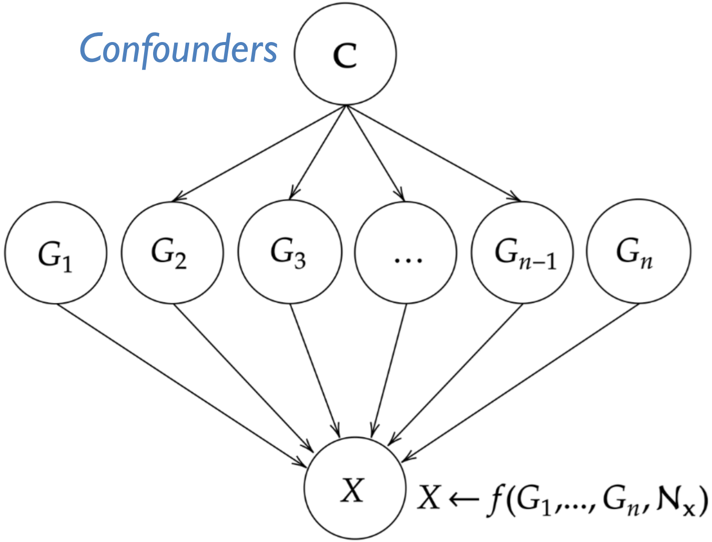
Causal Disentanglement

Our Work

Disentangled Causal Process

Confounders

Generative Factors



Causal model for X is disentangled

(iff)

it can be described by the SCM:

$$C_j \leftarrow \mathcal{N}_{C_j};\, j \in \{1,\ldots,l\}$$
$$G_i \leftarrow g_i(PA_i^C, \mathcal{N}_{G_i});\, i \in \{1,\ldots,n\}$$
$$X \leftarrow f(G_1,\ldots,G_n,\mathcal{N}_x)$$

f, g_i are independent causal mechanisms

Suter et al, Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness, ICML 2019

Evaluating Causal Disentanglement

Can Latent Variable Models (LVMs) learn to *causally* disentangle?

Metric 1: Unconfoundedness

- Encoder e of a LVM $\mathcal{M}$ (e, g, p_X) should learn the *mapping* from G_i to $\mathbf{Z}_I$ without any influence from C.

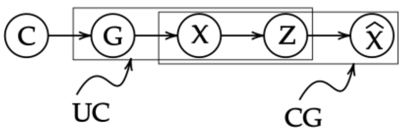
$$UC := 1 - \mathbb{E}_{x \sim p_X}\left[\frac{1}{S}\sum_{I,J}\frac{|\mathbf{Z}_I^x \cap \mathbf{Z}_J^x|}{|\mathbf{Z}_I^x \cup \mathbf{Z}_J^x|}\right]$$

Metric 2: Counterfactual Generativeness

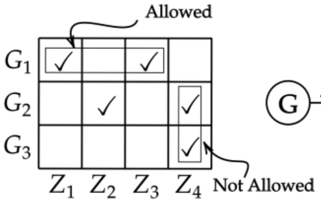
- If $\mathbf{Z}$ is unconfounded, the counterfactual of x w.r.t. G_i, x_I^{cf} can be generated by intervening on $\mathbf{Z}_I^x$.
- Any change in $\mathbf{Z}_{\setminus I}^x$, should have no influence on x_I^{cf} w.r.t. G_i.

$$CG := \mathbb{E}_I[|ACE_{\mathbf{Z}_I^X}^{X_I^{cf}} - ACE_{\mathbf{Z}_{\setminus I}^X}^{X_{\setminus I}^{cf}}|]$$

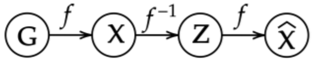
ACE = Average Causal Effect



UC & CG



UC



CG

Weakly Supervised Disentanglement

A Method

- Reconstruction *vs* Disentanglement!
- We use bounding box supervision for better trade-off
- We call our method *Semi-Supervised Factor-VAE with additional Bounding Box supervision* (*SS-FVAE-BB*).
- Augment Factor-VAE loss as:

$$\mathcal{L}_{SS-FVAE-BB} = \mathcal{L}_{(Factor-VAE)} + \lambda \sum_{i=1}^{L} ||x_i \odot w_i - \hat{x}_i \odot w_i||_2^2 \quad (4)$$

- $w_i \in \{0, 1\}^{320 \times 240 \times 3}$ is an indicator tensor with 1s in the region of the bounding box and 0s elsewhere.

Results

A Method

Model	*IRS*	*DCI* (D)	*UC* $\rho = 5$	*CG* $\rho = 5$	*UC* $\rho = 7$	*CG* $\rho = 7$
β-VAE	0.85	0.18	0.11	0.24	0.08	0.22
β-TCVAE	0.82	0.10	0.11	0.25	0.08	0.25
DIP-VAE	0.33	0.08	0.11	0.21	0.15	0.22
Factor-VAE	**0.88**	0.15	0.13	0.26	0.08	**0.28**
SS-β-VAE	0.74	**0.18**	0.11	**0.28**	0.08	0.19
SS-β-TCVAE	0.68	0.17	0.11	0.23	0.08	0.19
SS-DIP-VAE	0.35	0.08	0.11	0.22	0.15	0.22
SS-Factor-VAE	0.61	0.16	0.24	**0.28**	0.14	0.22
SS-FVAE-BB	0.61	0.13	**0.27**	**0.28**	**0.18**	**0.28**

- Low *UC*, *CG* scores indicate limitations in disentanglement achieved by SOTA models.
- SS-FVAE-BB achieves better *UC*, *CG* scores.

More information?

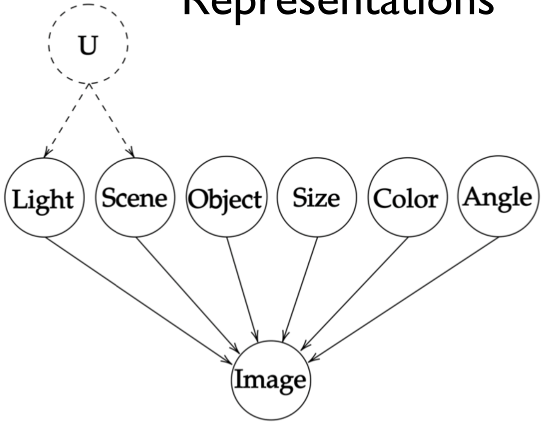
For more details:
https://arxiv.org/abs/2112.05746
https://github.com/causal-disentanglement/CANDLE

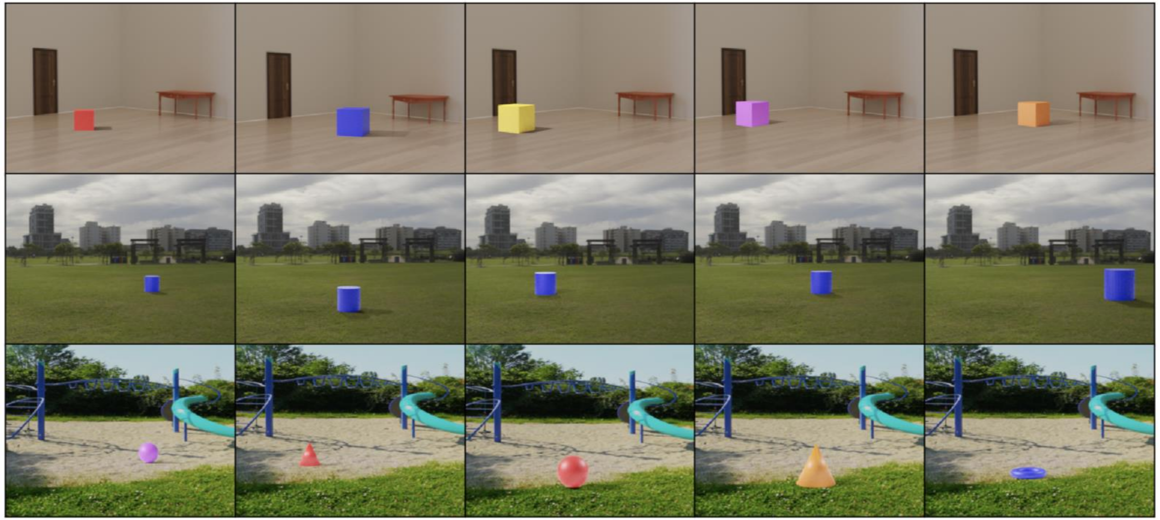
Our Other Ongoing Efforts in XAI

- Learning Causal Models on Latent Variables in Vision
- Concept-based Explanations in Vision
- Counterfactual Generation under Confounding
- Learning Disentangled Generative Processes and Mechanisms
- Causal Representation Learning
- ….

Need for Datasets/Benchmarks

CANDLE: An Image Dataset for Causal Analysis in Disentangled Representations





Best Paper Award, CVPR 2021 Workshop on Causality in Vision

https://github.com/causal-disentanglement/CANDLE

Open Problems and Challenges

- Is there a universal formalization for explainable ML?
- How to balance accuracy/performance vs interpretability tradeoff? Is interpretability always required?
- How to evaluate explainable systems?
- Who owns the explanation? Model or explanation methodology?
- How can connectionist and symbolic AI work together for 'logical' explanations?

Thank you!

Acknowledgements

Questions?

vineethnb@cse.iith.ac.in

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