Introduction

- Ante-hoc vs. post-hoc explanations
- Causal effects are reliable and human-centric
- Learn and explain causal effects in an ante-hoc manner
- Study various causal effects of input neurons on the output neurons[1]:
    - Average Controlled Direct Causal Effect (ACDE)
    - Average Natural Direct Causal Effect (ANDE)
    - Average Natural Indirect Causal Effect (AICE)
    - Average Total Causal Effect (ATCE)
- How to incorporate such causal effects in NNs?
- ACDE, ANDE, ATCE in CREDO (ICML 2022)
- AICE in AHCE (Under review)

[1] Judea Pearl. "Direct and indirect effects". In: *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*. 2001.

Direct and Indirect Causal Effects

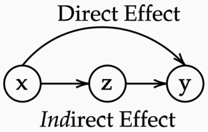
- Consider the causal effect of X on Y

$$ACE_x^y = \mathbb{E}[Y|do(X = x)] - \mathbb{E}[Y|do(X = x^*)]$$

- ACE_x^y is different from $\mathbb{E}[Y|X = x] - \mathbb{E}[Y|X = x^*]$

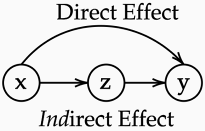
$$ACE_x^y = \mathbb{E}_W\mathbb{E}[Y|X = x, W = w)] - \mathbb{E}_W\mathbb{E}[Y|X = x^*, W = w]$$

- W is backdoor set and x^* is baseline intervention



- For direct causal effect, stop the influence through $X \to Z \to Y$
- For indirect causal effect, stop the influence flowing through $X \to Y$

Direct and Indirect Effects



- Direct Causal Effect

$$ADCE_X^Y = \mathbb{E}[Y|do(X = x, Z = Z_{x^*})] - \mathbb{E}[Y|do(X = x^*, Z = Z_{x^*})]$$

- Indirect Causal Effect

$$AICE_X^Y = \mathbb{E}[Y|do(X = x^*, Z = Z_x)] - \mathbb{E}[Y|do(X = x^*, Z = Z_{x^*})]$$

- Total Causal Effect

$$ATCE_X^Y = \mathbb{E}[Y|do(X = x, Z = Z_x)] - \mathbb{E}[Y|do(X = x^*, Z = Z_{x^*})]$$

Matching Learned Causal Effects of Neural Networks with Domain Priors (CREDO)

Abbavaram Gowtham Reddy*
Sai Srinivas Kancheti*
Vineeth N Balasubramanian
Amit Sharma

IIT Hyderabad
Lab1055 MLV Group @ IIT Hyderabad

ICML
International Conference On Machine Learning

Introduction

- Incorporate causal prior knowledge in NNs
- Priors in the form of (parametric) functional relationships
- Causal priors are a result of RCTs or come from domain knowledge
- Three kinds of priors motivated by 3 kinds of causal effects:
    - Average Controlled Direct Effect (ACDE)
    - Average Natural Direct Effect (ANDE)
    - Average Total Causal Effect (ATCE)
- We incorporate them in NNs by gradient-based regularization.

Notations & Background

- We view a feed forward NN f as a structural causal model
- Neurons represent variables and edges represent causal relationships
- Mrginalize over hidden layers of a neuron and consider only input and output layers.
- Let $\mathcal{G}$ be the causal graph of the SCM of f in which
  - T is the treatment variable
  - $\hat{Y}$ is the outcome variable
  - Z is the set of variable that lie in a directed path from T to $\hat{Y}$ (in the NN causal graph).
  - W is the set of remaining variables
  - We denote $\hat{Y}|do(T = t)$ as $\hat{Y}_t$

Different Causal Effects

- A trained NN learns some causal relationships between the inputs and the outputs
- Following Pearl[2], we define various causal effects of the feature T on $\hat{Y}$ learned by NN SCM
- First we define the ACDE in NNs and show its identifiability
- Please refer to our paper[3] for regularizing and explaining ANDE, ATCE.
- Controlled direct effect is slightly different from natural direct effect.
- In ACDE, intervention on Z, W is fixed instead of deriving from x^*.

---

[2] Judea Pearl. "Direct and indirect effects". In: *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*. 2001.

[3] Sai Srinivas Kancheti et al. "Matching Learned Causal Effects of Neural Networks with Domain Priors". In: *ICML*. PMLR. 2022.

Different Causal Effects

Average Controlled Direct Effect (ACDE) in NNs

Average Controlled Direct Effect (*NN-ACDE*) measures the average causal effect of T on $\hat{Y}$ when all parents of $\hat{Y}$ except T (Z, W in this case) are intervened to pre-defined control values (i.e., $do(Z = z, W = w)$).

$NN\text{-}ACDE_t^{\hat{Y}}(z, w) := \mathbb{E}_U[\hat{Y}_{t,z,w}] - \mathbb{E}_U[\hat{Y}_{t^*,z,w}] = \hat{Y}_{t,z,w} - \hat{Y}_{t^*,z,w}$.

- Priors are expressed only in terms of T and Y
- A modified definition for *NN-ACDE* that marginalizes over $\{Z, W\}$.

Our version of *NN-ACDE* is hence:

$$NN\text{-}ACDE_t^{\hat{Y}} := \mathbb{E}_{Z,W,U}[\hat{Y}_{t,Z,W}] - \mathbb{E}_{Z,W,U}[\hat{Y}_{t^*,Z,W}]$$

Similarly, we define *NN-ANDE* and *NN-ATCE* in NNs.

Identifying Causal Effects

Identifying ACDE in NNs

$$
\begin{aligned}
ACDE_t^{\hat{Y}} &= \mathbb{E}_{Z,W,U}[\hat{Y}_{t,Z,W}] - \mathbb{E}_{Z,W,U}[\hat{Y}_{t^*,Z,W}] \ (\textit{Definition}) \\
&= \mathbb{E}_{Z,W}[\hat{Y}_{t,Z,W}] - \mathbb{E}_{Z,W}[\hat{Y}_{t^*,Z,W}] \ (\textit{NN is deterministic}) \\
&= \mathbb{E}_{Z,W}[\hat{Y}|t, Z, W] - \mathbb{E}_{Z,W}[\hat{Y}|t^*, Z, W] \ (\textit{Unconfoundedness})
\end{aligned}
$$

- The ACDE can be computed empirically by sampling Z, W (covariates other than T) from training data, and computing $\hat{Y}$ via forward pass
- Similarly, we prove the identifiability of *NN-ANDE* and *NN-ATCE* in NNs

Regularizing Causal Effects

- Match the causal effects learned by the NN to the true causal effects
- We enforce this by gradient matching
- The gradient of the provided causal domain prior is matched with the gradient of the NN's learned causal effect
- Gradient matching for ACDE is straightforward
- Gradient matching for ANDE should be done at a specific point derived from t^* i.e., (t, Z_{t^*}, W)
- We match the total derivative to regularize ATCE

Regularizing Causal Effects

Regularizing ACDE in NNs

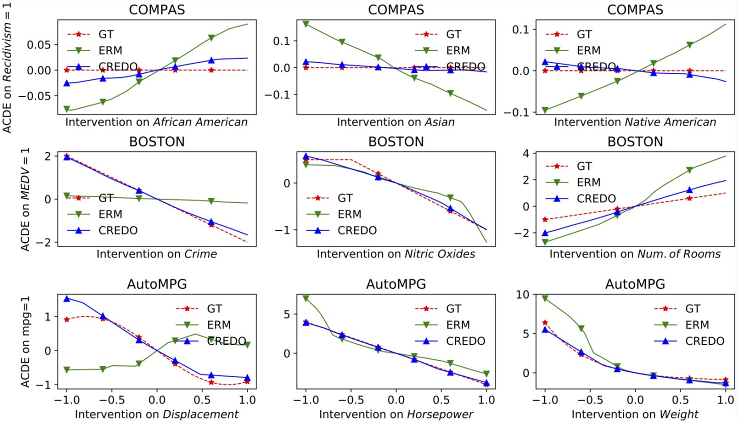
$$\frac{\partial ACDE_t^{\hat{Y}}}{\partial t} = \frac{\partial[\mathbb{E}_{Z,W}[\hat{Y}|t,Z,W] - \mathbb{E}_{Z,W}[\hat{Y}|t^*,Z,W]]}{\partial t}$$
$$= \frac{\partial[\mathbb{E}_{Z,W}[\hat{Y}|t,Z,W]]}{\partial t} \; (t^* \text{ is a constant})$$
$$= \mathbb{E}_{Z,W}\left[\frac{\partial[\hat{Y}(t,Z,W)]}{\partial t}\right] \; (\text{exchange } \mathbb{E} \text{ and } \frac{\partial}{\partial t})$$

Regularizer

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

Similarly, we regularize ANDE and ATCE in NNs

Results



ACDE plots. The blue curves closely matches the domain priors (red curves), indicatint that CREDO learns the desired causal effects

Thank You!