Matching Learned Causal Effects of Neural Networks with Domain Priors

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

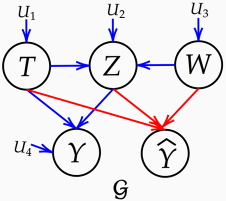
Introduction

- We consider causal domain priors in the form of (parametric) functional relationships between inputs and outputs
- Domain priors often come as a result of RCTs or from domain knowledge
- We consider 3 kinds of domain priors motivated from 3 kinds of causal effects defined by Pearl[1]:
    - Average Controlled Direct Effect (ACDE)
    - Average Natural Direct Effect (ANDE)
    - Average Total Causal Effect (ATCE)
- If we know such priors, we incorporate them in neural networks (NNs) by regularization.

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

Notations & Background

- We view a feed forward NN f as a structural causal model
- Neurons represent variables and edges represent causal relationships among variables.
- W.l.o.g, we marginalize 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

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

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
- We propose 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

- We would like to match the causal effects learned by the NN to the true causal effects which are provided to us in the form of causal domain priors
- 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

Regularizing Causal Effects

Regularizing ACDE in NNs

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

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

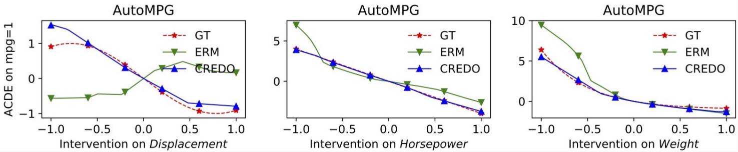
$$= \mathbb{E}_{Z,W}\left[\frac{\partial[\hat{Y}(t, Z, W)]}{\partial t}\right] \ (exchange \ \mathbb{E} \ and \ \frac{\partial}{\partial t})$$

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 of AutoMPG dataset

The blue curve closely matches the domain prior (red curve), which indicates that CREDO (the causally regularized NN) learns the desired causal effects

Thank You!