CIS II Paper Review

Author:

Chang Yan

cyan13@jhu.edu

Paper:

Deep Structural Causal Models for Tractable Counterfactual Inference

Pawlowski, N., Castro, D. C., & Glocker, B. (2020). Deep structural causal models for tractable counterfactual inference. arXiv preprint arXiv:2006.06485.

Paper selection reason:

1. Deep Structural Causal Models are novel inventions aiming to improve the current unsatisfactory performance of most deep neural networks on image prediction.
2. One of their experiments was to predict MRI scans, which has some similarity to our task of hemothorax classification on CT scans, so we might be able to adapt some of their means.
3. Their project is open-source and uses PyTorch and Pyro based algorithms, compatible with our project structure.
4. The Deep Structural Causal Models focuses on the causality of association, interventions and counterfactuals of events, which provides the connection between parameters is deep networks and the actual biological events happening, which might help us improve our classification as hemothorax is a biological event with strong causality.
5. The paper was first published in less than 1 years ago, reflecting novel and cutting-edge developments of deep learning.

CIS II Project Summary:

We are developing deep-learning based algorithms to perform 3D segmentation on CT scans, and predict hemothorax volume as voxel count accordingly. Also, the 3D segmentation result helps human operator to assess the quality of volume prediction.

Authors' problems and goals:

Problems of current deep learning networks:

1. DL is known to be susceptible to learning spurious correlations.
2. DL tend to amplify biases.
3. DL is exceptionally vulnerable to changes in the input distribution.

Problems of current Structural Causal Models:

1. SCMs are typically employed with simple linear mechanisms
2. works well for scalar variables and can be useful for decision making, but is not flexible enough to model higher-dimensional data such as images

Goals:

1. Develop a general framework for building structural causal models (SCMs) with deep learning components, called DSCMs, to solve the problems above.
2. Employ normalizing flows and variational inference to enable tractable inference of exogenous

noise variables—a crucial step for counterfactual inference that is missing from existing deep causal learning methods.

Significance of study:

1. Causal DL models could be capable of learning relationships from complex high-dimensional data and of providing answers to interventional and counterfactual questions.
2. By explicitly modelling causal relationships and acknowledging the difference between causation and correlation, causality becomes a natural field of study for improving the transparency, fairness, and robustness of DL based systems
3. The tractable inference of deep counterfactuals enables novel research avenues that aim to study causal reasoning on a per instance rather than population level, which could lead to advances in personalized medicine as well as in decision-support systems, more generally.

Background information:

1. Pearl's ladder of causation

- **Association**
  describes reasoning about passively observed data. This level deals with correlations in the data and questions of the type "What are the odds that I observe. . . ?"
- **Intervention**
  concerns interactions with the environment. It requires knowledge beyond just observations, as it relies on structural assumptions about the underlying data-generating process. Characteristic questions is "What happens if I do. . . ?"
- **Counterfactuals**
  deal with retrospective hypothetical scenarios. Counterfactual is the generative processes to imagine alternative outcomes for individual data points, answering "What if I had done A instead of B?"

2. Structural causal models and how them fulfill the ladder of causation

$$\mathfrak{G} := (\mathbf{S}, P(\boldsymbol{\epsilon}))$$

$$\mathbf{S} = (f_1, \ldots, f_K)$$

f_k is their structural assignments.

$$x_k := f_k(\epsilon_k; \mathbf{pa}_k)$$

x_k is the events, ε_k is the exogenous noise, pa_k is the set of direct causes of x_k,

$$P(\boldsymbol{\epsilon}) = \prod_{k=1}^{K} P(\epsilon_k)$$

P(ε) is the joint distribution over mutually independent exogenous noise variables.

- **Association**
  Embedded in this model
- **Intervention**
  do(x_k := a). disconnect x_k with its parents and change structural assignment $f_{k.}$ Possible of changing both S and P(ε).
- **Counterfactuals**
  hypothetical retrospective interventions: 'What would x_i have been if x_j were different, given

that we observed x?' Only change S, not P(ε)

Do mathematically in 3 steps:

- **Abduction:**

  Predict the 'state of the world' (the exogenous noise, ε) that is compatible with the observations, x, i.e. infer P(ε|x)

- **Action:**

  Perform an intervention (e.g. do(x_k := x_k')) corresponding to the desired manipulation, resulting in a modified SCM G' = (S', P(ε|x))

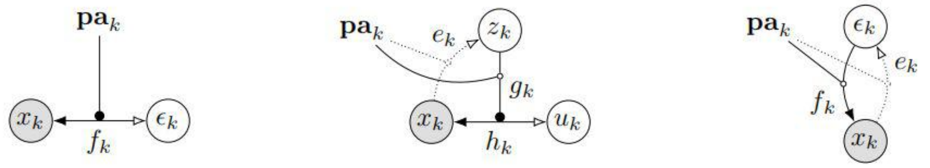
- **Prediction:**

  Compute the quantity of interest based on the distribution entailed by the new counterfactual SCM as P(x).

Authors' work and implementations:

They use recent advances in normalizing flows and variational inference to model mechanisms for composable DSCMs that enable tractable counterfactual inference.

As the mathematical basis is too complicated and requires professional on statistics, I discuss them using the causal graphs the authors provided here.



(a) Invertible explicit likelihood (b) Amortised explicit likelihood (c) Amortised implicit likelihood

Here, white arrows indicates abductive direction, and black arrows indicates generative direction. Dotted lines are amortized variational approximation. f_k is the forward model, e_k is an encoder that amortizes abduction in non-invertible mechanisms, g_k is a 'high-level' non-invertible branch (e. g. a probabilistic decoder), and hk is a 'low-level' invertible mapping (e.g. reparametrization).

(a) Invertible explicit likelihood: a flow-based generative model constructed by a sequence of invertible transformations. It explicitly learns the data distributions:

$$x_i := f_i(\epsilon_i; \mathbf{pa}_i), \qquad p(x_i \,|\, \mathbf{pa}_i) = p(\epsilon_i) \cdot |\det \nabla_{\epsilon_i} f_i(\epsilon_i; \mathbf{pa}_i)|^{-1}\big|_{\epsilon_i = f_i^{-1}(x_i; \mathbf{pa}_i)}$$

(b) Amortized explicit likelihood: The invertible model has heavy computational requirements when modelling high-dimensional observations, so the authors proposes to separate the assignment f_k into a 'low-level', invertible component h_k (often a convolutional neural net work) and a 'high-level', non-invertible part g_k (often a probabilistic decoder), with corresponding noise decomposition ε_k (an encoder for amortized variational approximation of abduction):

$$x_k := f_k(\epsilon_k; \mathbf{pa}_k) = h_k(u_k; g_k(z_k; \mathbf{pa}_k), \mathbf{pa}_k), \qquad P(\epsilon_k) = P(u_k)P(z_k)\,.$$

$$p(x_k \,|\, z_k, \mathbf{pa}_k) = p(u_k) \cdot |\det \nabla_{u_k} h_k(u_k; g_k(z_k, \mathbf{pa}_k), \mathbf{pa}_k)|^{-1}\big|_{u_k = h_k^{-1}(x_k; g_k(z_k, \mathbf{pa}_k), \mathbf{pa}_k)}$$

(c) Amortized implicit likelihood: Casual graph provided only for completeness, not implemented.

Deep counterfactual inference algorithm they implemented:

- **Abduction:**

Use the trained encoder e_j to approximate ε_j

$$\epsilon_j \approx e_j(x_j; \mathbf{pa}_j)$$

And calculate approximated P(ε|x, pa_k)

$$\begin{aligned} P_{\mathfrak{G}}(\epsilon_k \,|\, x_k, \mathbf{pa}_k) &= P_{\mathfrak{G}}(z_k \,|\, x_k, \mathbf{pa}_k)\, P_{\mathfrak{G}}(u_k \,|\, z_k, x_k, \mathbf{pa}_k) \\ &\approx Q(z_k \,|\, e_k(x_k; \mathbf{pa}_k))\, \delta_{h_k^{-1}(x_k; g_k(z_k; \mathbf{pa}_k), \mathbf{pa}_k)}(u_k) \end{aligned}$$

- **Action:**

Replace x_k by either a constant x_k := x_k' or by surrogate mechanism x_k := f_k'(ε_k, pa_k) resulting in a modified SCM G' = (S', P(ε|x))

- **Prediction:**

First approximate the counterfactual distribution using Monte Carlo method:

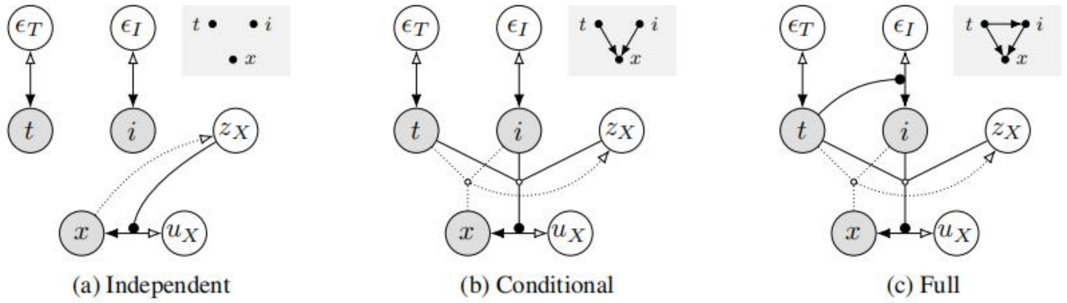
$$\begin{aligned} z_k^{(s)} &\sim Q(z_k \,|\, e_k(x_k; \mathbf{pa}_k)) \\ u_k^{(s)} &= h_k^{-1}(x_k; g_k(z_k^{(s)}; \mathbf{pa}_k), \mathbf{pa}_k) \\ \widetilde{x}_k^{(s)} &= \widetilde{h}_k(u_k^{(s)}; \widetilde{g}_k(z_k^{(s)}; \widetilde{\mathbf{pa}}_k), \widetilde{\mathbf{pa}}_k)\,. \end{aligned}$$

Sample from the distribution using uncorrelated Gaussian decoder for images:

$$\widetilde{x}_k^{(s)} = x_k + [\mu(z_k^{(s)}; \widetilde{\mathbf{pa}}_k) - \mu(z_k^{(s)}; \mathbf{pa}_k)]\,.$$

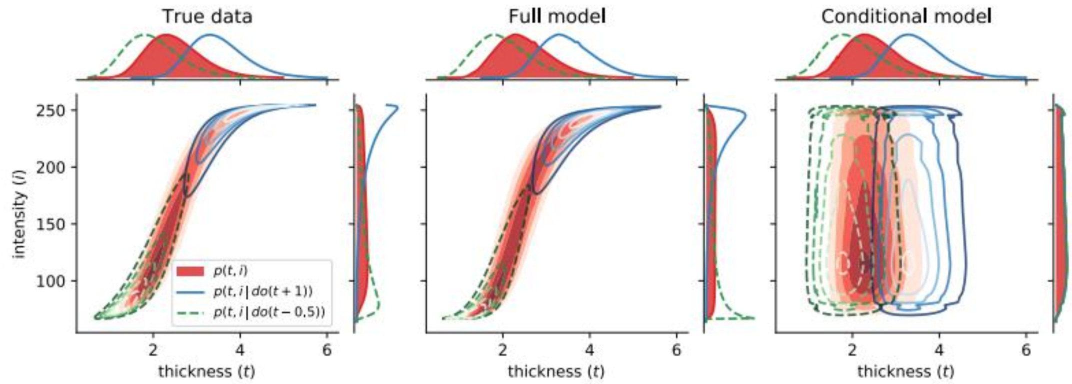
Authors' experiments and results:

Case Study I: Morpho-MNIST:

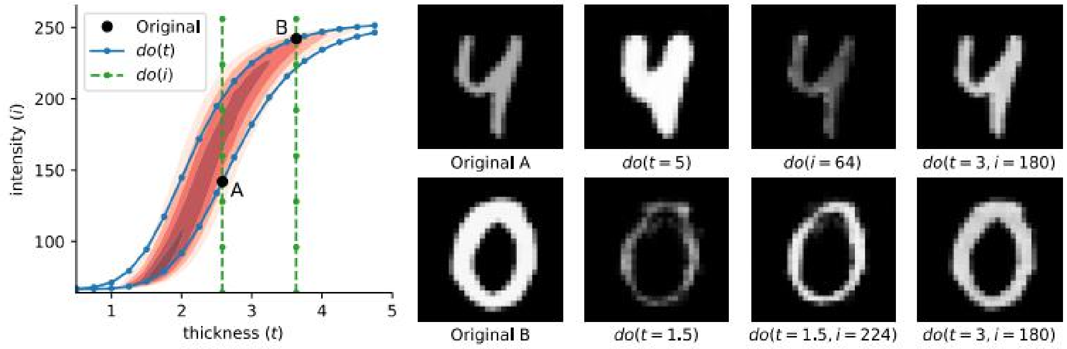


(a) Independent (b) Conditional (c) Full

They test three different models in a synthetic dataset based on MNIST digits, where they defined stroke thickness to cause the brightness of each digit: thicker digits are thicker digits are brighter whereas thinner digits are dimmer. The result is as follow:



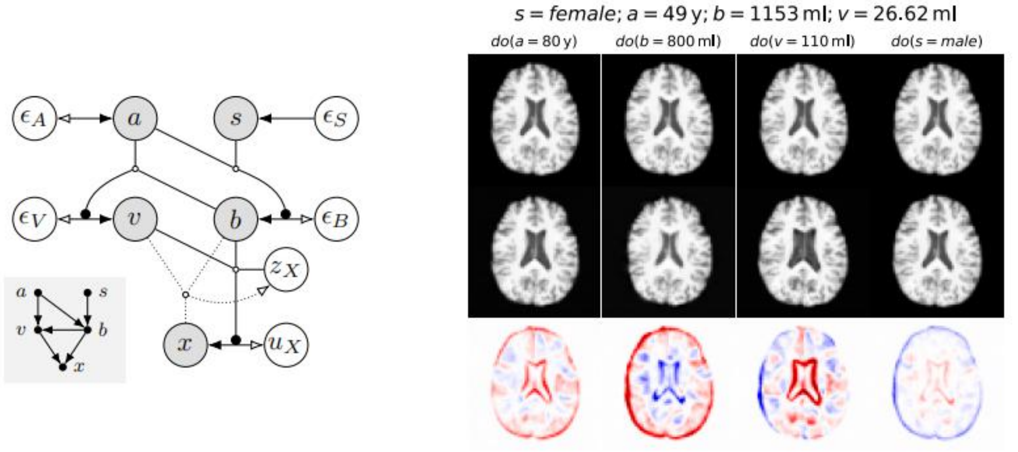
We can see that the full model is very close to the true data, while the conditional model yield inaccurate result due to the lack of the casual relationship between thickness and intensity.



We can see that modifying intensity does not change thickness at all, but change thickness also changes intensity. This is the indication that the authors have correctly modeled a causal relationship where thickness change causes intensity change, but intensity change does not cause thickness change. The result shows that their model is capable of generating convincing counterfactuals that preserve the digit identity while changing thickness and intensity consistently.

Case Study II: Brain MRI

The author models the change of MRI based on patients' sex, age, brain volumes and ventricle flows. The DSCM structure is shown as follow:



(a) Computational graph

(b) Original image, counterfactuals, and difference maps

In their model, age and sex affects brain volume, age and brain volume affects ventricle flow, and ventricle flow and brain volume together show on the image. Also, from the image we can approximate abduction to ventricle flow and brain volume.

The difference maps show plausible counterfactual changes: increasing age causes slightly larger ventricles while decreasing the overall brain volume (first column). In contrast, directly changing brain volume has an opposite effect on the ventricles compared to changing age (second column). Intervening on ventricle volume has a much more localized effect (third column), while intervening on the categorical variable of biological sex has smaller yet more diffuse effects.

Importance and relevance to me:

1. It provides me with a deep insight with SCMs and how they can be combined with and improve DL.
2. The causal inference they developed could possibility be added to our model to improve explainability and assess confidence.

Good points they did:

1. They proposed a novel process of integrating SCM and DL with sufficient mathematical basis, and also proposed specific mathematical ways to model counterfactual inference, which other studies fail to achieve.
2. They used casual graphs to explain their model designs, which are clear, informative and easy to understand.
3. They provided a set of pictures to show the affect of changing each variables, which are very intuitive.

Criticisms

1. Although they properly defined the mathematical basis, they did not talk much into the actual structure of the neural networks they design. However, the deep network structure itself is also crucial to the performance.
2. What’s worse, their code is completely undocumented, and badly structured. It is really hard to replicate their work.
3. They actually used over 10 different decoders in their code, but in their paper they only talked about the Gaussian decoder, not the others.
4. When assessing the casual relationship in the brain MRI model, they have 4 variables and only change one at a time. However, I think it is better to also explore the change of more than one variables to verify the correct casual relationships.
5. On experiment one, when the i variable is neither direct nor indirect cause of t, changing i does not change t at all. However, on experiment two, the s variable is also neither direct nor indirect cause of v, but apparently changing s has some affect on v. They did not say anything about why there is a difference.
6. Most of their assessment of causality in results are based on qualitative analysis, not quantitative. It would be better if they can say something mathematically about how accurate their prediction is. They only showed that the DSCM can model causality, but not tested if those modeled relationships are actually correct and how they compared to ground truth.

Possible next steps:

1. Develop a way to use DSCM to discover implicit causalities, instead of only modelling assumed causalities.
2. Add a more robust and mathematical way to assess the quality of prediction.