WhatIfGAN Performance on image mediator experiment

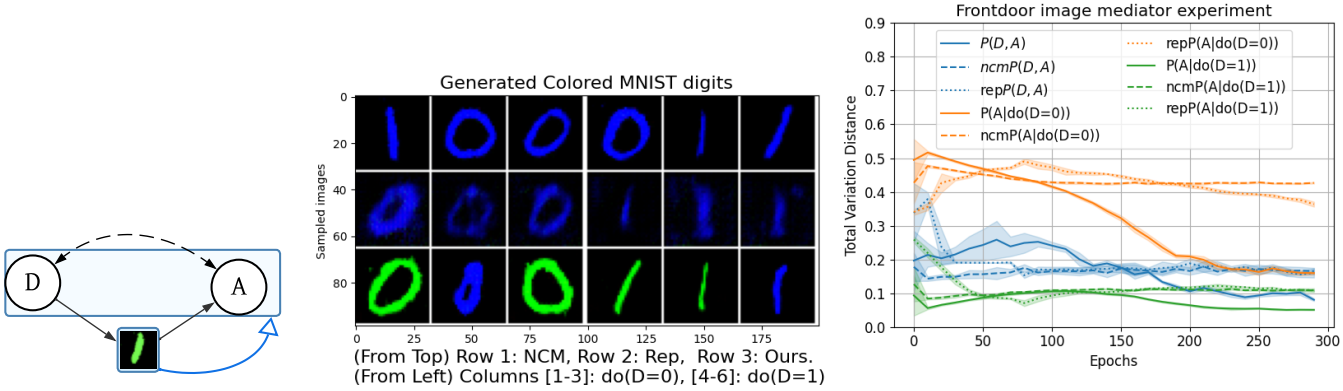
rahman89

July 2023

To illustrate WhatIfGAN performance in learning the joint distribution involving both low and high-dimensional variables, we ran an experiment on the front-door graph in Figure 1a. We constructed a synthetic SCM where a variable U affects both D and A binary variables but is kept hidden in the dataset to make it act like a confounder. Image variable I contains the digit value of D, and A is some attribute of I. Suppose we are given a dataset sampled from $P(D, A, I)$ distribution. Our goal is to estimate the causal effect of D on variable A. To measure the ground truth causal effect, we can use the backdoor criterion (Pearl, 1993), $P(A|do(D)) = \sum_U P(A|D, U)P(U)$ since we have access to U in the true SCM. In the observational dataset, $P(A|do(D))$ is identifiable with the front-door criterion (Pearl, 2009): $P(A|do(D)) = \sum_I P(I|D) \sum_{D'} P(A|D', I)P(D')$.

If we can train all mechanisms in the DCM to match $P(D, A, I)$, we can produce correct samples from $P(A|do(D))$. Now, the identification algorithm can not be applied since it requires image distribution. Even GAN convergence becomes difficult using the joint distribution loss since the losses generated by low and high dimensional variables are not easily comparable and it is non-trivial to find a correct re-weighting of such different loss terms. To the best of our knowledge, no current causal effect estimation algorithm can address this problem since there is no estimator that does not contain explicit image distribution, which is practically impossible to estimate. To deal with this problem, we map samples of I to a low-dimensional representation, RI with a trained encoder and match $P(D, A, RI)$ instead of the joint $P(D, A, I)$. We construct the rest of the WhatIfGAN architecture with a neural network having fully-connected layers to produce D, a deep convolution GAN to generate images and a classifier to classify MNIST images into variable A such that D and A are confounded. Now, for this graph, the corresponding $\mathcal{H}$-graph is $[I] \rightarrow [D, A]$. We can train $\mathbb{G}_I$ by matching $P(I|D)$. Instead of training $\mathbb{G}_I$, we can also employ a pre-trained generative model that takes digits D as input and produces a colored MNIST image showing D digit in it. Next, to train $\mathbb{G}_D$ and $\mathbb{G}_A$, we should match the joint distribution $P(D, A, I)$ since $\{I\}$ is ancestor set $\mathcal{A}$ for c-component $\{D, A\}$.

In Figure 1b and 1c, we compare our method with Xia et al. (2023): NCM and a version of our method: WhatIfGAN-rep that does not use modular training. NCM implementation cannot be directly used for high-dimensional data. Thus we implemented their approach on our architectures. Since NCM trains all mechanisms with the same loss function calculated from both low and high-dimensional samples, it learns marginal distribution $P(I)$ (Figure 1b row-1) but does not converge to match the joint $P(D, A, I)$ (Figure 1c dashed-lines). As a result NCM produces good quality image but not consistent with $do(D)$ intervention. WhatIfGAN-rep uses a low-dim representation of images: RI and matches the joint distribution $P(D, A, RI)$ as a proxy to $P(D, A, I)$ without modularization. We observe WhatIfGAN-rep to converge (Figure 1c dotted-lines) slower compared to the original WhatIfGAN and produce low-quality images (Figure 1b row-2). Thus, WhatIfGAN-rep produces consistent but low quality images for $do(D)$ intervention. Finally, WhatIfGAN modular training matches $P(D, A, RI)$ and converges faster (Figure 1c solid-lines) for $P(D, A), P(A|do(D))$ and produces high-quality $P(I|do(D))$ images (Figure 1b row-3). Therefore, produced images are high quality and consistent with $do(D)$ intervention.



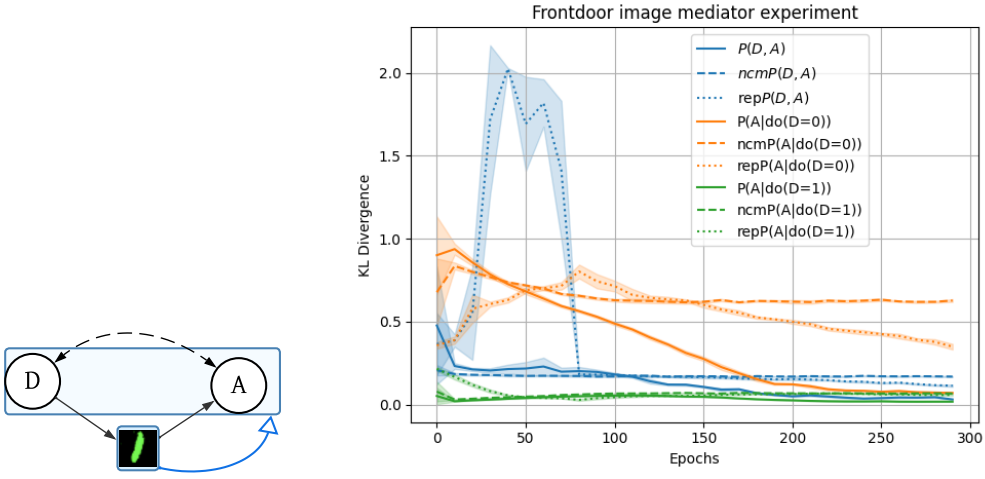
(a) Frontdoor causal graph w/ image mediator

(b) MNIST Image generation comparison

(c) Training converges matching $P(D, A)$ and $P(A|do(D))$

Figure 1: Modular Training on frontdoor causal graph with training order: $\{I\} \rightarrow \{D, A\}$. NCM produces good quality image but not consistent with $do(D)$ intervention. WhatIfGAN-rep produces consistent but low quality images. WhatIfGAN produces both consistent and high quality images.

More details about image mediator experiment



(a) Frontdoor causal graph w/ image mediator

(b) Training converges matching $P(D, A)$ and $P(A|do(D))$

Figure 2: Modular Training on frontdoor causal graph with training order: $\{I\} \rightarrow \{D, A\}$

We have domain $D = [0, 1]$, Image size=$3 \times 32 \times 32$ and $A = [0, 1, 2]$. Let U_0, e_1, e_2, e_3 are randomly generate exogenous noise. $D = U_0 + e_1, Image = f_2(D, e_2), A = f_3(Image, e_3, U_0)$. f_2 is a function which takes D and e_2 as input and produces different colored images showing D digit in it. f_3 is a classifier with random weights that takes U_0 and $Image$ as input and produces A such a way that $|P(A|do(D = 0)) - P(A|D = 0)|, |P(A|do(D = 1)) - P(A|do(D = 0))|and|P(A|D = 1) - P(A|D = 0)|$ is high. We calculate ground truth of P(C|do(D)) with backdoor criterion.

$$P(A|do(D)) = \sum_{U_0} P(A|D, U_0)P(D|U_0)$$

WhatIfGAN samples from $P(A|do(D))$ after training. The query is identifiable with frontdoor criterion when U_0 is unobserved. Image is a mediator here.

$$P(A|do(D)) = \sum_{Image} P(Image|D) \sum_{D'} P(A|D', Image)P(D')$$

This inference is not possible with identification algorithm. WhatIfGAN can achieve that by producing Image samples instead of learning its distribution. We use a lower-dimensional representation RI of Image variable and match $P(D, RI, A)$ instead of $P(D, Image, A)$. Samples from $P(Image|do(D = 1))$

References

Pearl, J. (1993). [bayesian analysis in expert systems]: comment: graphical models, causality and intervention. *Statistical Science*, 8(3):266–269.

Pearl, J. (2009). *Causality*. Cambridge university press.

Xia, K. M., Pan, Y., and Bareinboim, E. (2023). Neural causal models for counterfactual identification and estimation. In *The Eleventh International Conference on Learning Representations*.