On Causally Disentangled Representations

Abbavaram Gowtham Reddy, Benin Godfrey L, Vineeth N Balasubramanian

Department of Computer Science and Engineering, Indian Institute of Technology Hyderabad, India.

AAAI-22

IITH-CANDLE

Contributions

- We study disentanglement in Latent Variable Models (LVM) (e.g., $\beta-$VAE) from a causal perspective.
- We study interventional and counterfactual goodness in LVMs.
- We consider generative processes where generative factors are either independent or can be potentially confounded.
- We present two evaluation metrics that are consequences of the properties of causally disentangled LVMs.
- We introduce a new image dataset, CANDLE, that includes known causal generative factors as well as confounders to study and improve LVMs from a causal perspective.
- We perform empirical studies on well-known LVMs and analyze their performance. We also show how a small degree of weak supervision can improve causal disentanglement.

Introduction

- We study 2-level causal generative processes of the form shown in Figure 1.
- We look at the essential properties of causal disentanglement and propose evaluation metrics using the first principles of causality.
- We introduce a new realistic image dataset, CANDLE, whose generation follows a graph similar to Figure 1.

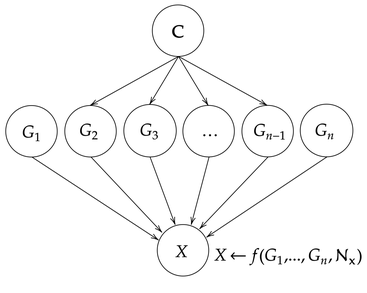

Figure 1:Data Generating Mechanism

Disentanglement in LVMs

G_i is the i^{th} generative factor, G_{ik} is the k^{th} value taken by G_i, $\mathbf{Z}$ is the set of latents in an LVM, $M = |\mathbf{Z}|, N = |G|$. $I \subset \{1, 2, \ldots, M\}$, $\mathbb{D} = \{x_i\}_{i=1}^{L}$ is dataset of images. Two properties should be satisfied by any disentangled LVM are:

Property 1: If an LVM disentangles a causal process, encoder learns a latent space $\mathbf{Z}$ such that each G_i is mapped to a unique $\mathbf{Z}_I$.

Property 2: In an LVM, that disentangles a causal process, the only causal features of $\hat{x}$ (reconstruction of x) w.r.t. G_i is $\mathbf{Z}_I \ \ \forall i$.

Unconfoundedness (UC)

From Property 1, if an LVM is able to map each G_i to a unique $\mathbf{Z}_I$, we say that $\mathbf{Z}$ is unconfounded.

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

- We use IRS score to find $\mathbf{Z}_I^x$ corresponding to G_i

Counterfactual Generativeness (CG)

From Property 2, when $\mathbf{Z}$ is unconfounded, a counterfactual instance 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 relies on the ACE of $\mathbf{Z}$ on $\hat{x}$
- $CG := \mathbb{E}_I[|ACE_{\mathbf{Z}_I}^{X_I^{cf}} - ACE_{\mathbf{Z}_{\setminus I}}^{X_{\setminus I}^{cf}}|]$
- Approximating ACE with average of ICEs

$$CG \approx \frac{1}{N}\frac{1}{L}[|ICE_{\mathbf{Z}_I^x}^{x_I^{cf}} - ICE_{\mathbf{Z}_{\setminus I}^x}^{x_{\setminus I}^{cf}}|]$$

$$ICE_{\mathbf{Z}_I^x}^{x_I^{cf}} = |P(G_{ik}|x_I^{cf}, do(\mathbf{Z}_I = \mathbf{Z}_I^x)) - P(G_{ik}|x_I^{cf}, do(\mathbf{Z}_I = baseline(\mathbf{Z}_I))|$$

- We choose $baseline(\mathbf{Z}_I)$ as the latent values that are maximally deviated from from current latent values $\mathbf{Z}_I^x$ of an image x(taken over dataset)
- UC and CG are equal to 1 for perfect disentanglement and 0 for no disentanglement

CANDLE , A New Dataset For Causal Disentanglement

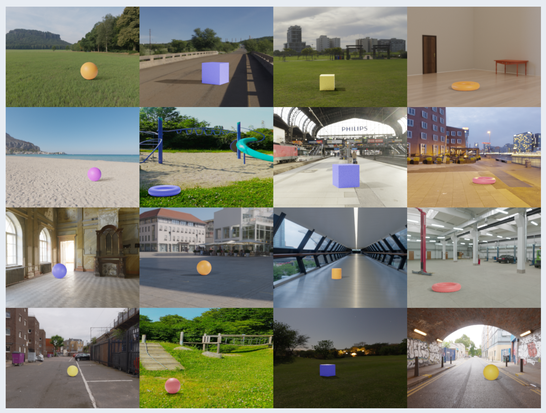

Figure 2:Sample Images from CANDLE dataset

- Background images in CANDLE are 4000×2000 resolution real HDRI images.
- Foreground objects are synthetic and placed on the floor with no occlusion.
- Objects are small for semantic correctness in relation to the background.
- Artificial light source is added which casts shadows in 3 positions: left, middle, and right.
- Meta data is provided in JSON format.
- CANDLE exhibits both observed and unobserved confounding in its images.

Weak Supervision

Using bounding box supervision in metadata, we propose an improvement, SS-FVAE-BB, over SOTA methods by augmenting the loss function of Factor-VAE to pay more attention to foreground objects.

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

where $w_i \in \{0,1\}^{320\times240\times3}$ is an indicator tensor with 1s in the region of the bounding box and 0s elsewhere, λ is a hyperparameter and $\odot$ is the Hadamard (elementwise) product.

Results

- We study β-VAE, DIP-VAE, Factor-VAE and β-TCVAE and corresponding semi supervised models. Models are compared with IRS, $DCI(D)$, UC and CG metrics.
- *SS-* prefix refers to the 'Semi-Supervised' variants. Semi supervision is provided in terms of labels to 10% of data points.
- SS-FVAE-BB achieves better UC, CG scores. Low UC, CG score indicate the limitations of SOTA methods in achieving causal disentanglement.
- ρ is the number of latent dimensions that we choose to attribute for each generative factor and the table below shows the results on CANDLE.

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**

References

- Pearl, Judea. Causality. Cambridge university press, 2009.
- Suter, R., Miladinovic, D., Schölkopf, B., Bauer, S. (2019, May). Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. In ICML (pp. 6056-6065). PMLR

Contact Information

- cs19resch11002@iith.ac.in
- benin.godfrey@cse.iith.ac.in
- vineethnb@iith.ac.in