CausalVAE: Disentangled Representation Learning via Neural Structural Causal Models

Mengyue Yang1,2 , Furui Liu1,*, Zhitang Chen1 , Xinwei Shen3 , Jianye Hao1 , Jun Wang2 1 Noah's Ark Lab, Huawei, Shenzhen, China 2 University College London, London, United Kingdom 3 The Hong Kong University of Science and Technology, Hong Kong, China
CVPR 2021

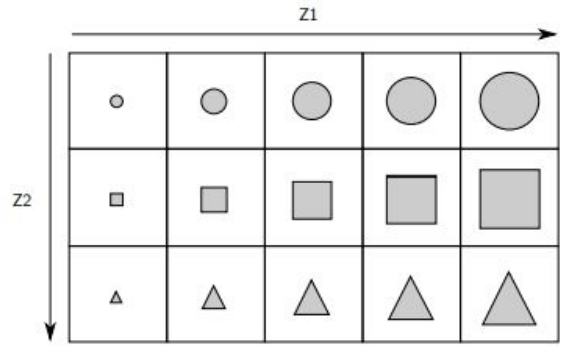
Nikoo Naghavian
Winter 2023

outline

- Introduction
- Model structure of causalVAE
- Learning strategy
- Experiments
- Methods comparison
- Conclusion
- Code explanation

Disentangled Representation

A disentangled representation can be defined as one where single latent units are sensitive to changes in single (independent) generative (ground-truth) factors, while being relatively invariant to changes in other factors



Variational Autoencoder (VAE)

- Setting:
  - Let $\boldsymbol{z}$ denote the concepts representation, $\boldsymbol{x}$ be observations
  - Using marginal distribution $p(\boldsymbol{x})$ to get joint distribution $p(\boldsymbol{x}, \boldsymbol{z})$
- Model:
  - Encoder: inference latent code $\boldsymbol{z}$
  - Decoder: generate/reconstruct $\boldsymbol{x}$ from $\boldsymbol{z}$

- Objective:
  - Maximize Evidence lower bound
    - $\mathbb{E}_{p_\theta(\boldsymbol{z})}[p_\theta(\boldsymbol{x})] \geq ELBO = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x})] - D_{kl}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}))$
  - $\mathbb{E}_{p_\theta(\boldsymbol{z})}[p_\theta(\boldsymbol{x}|\boldsymbol{z})]$: Reconstruct $\boldsymbol{x}$ from $\boldsymbol{z}$ in generate process
  - $D_{kl}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p_\theta(\boldsymbol{z}))$: Introduce posterior $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ and prior $p_\theta(\boldsymbol{z})$ (Multvariate Gaussian) in inference process

Problems in traditional disentanglement works

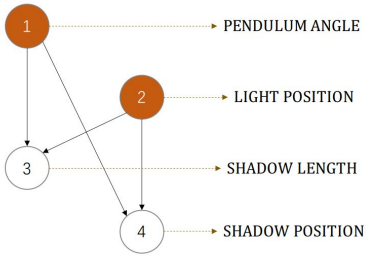
- The concepts are causally related.
- Unsupervised process could not guarantee the learned representation is identifiable.

Swing Pendulum

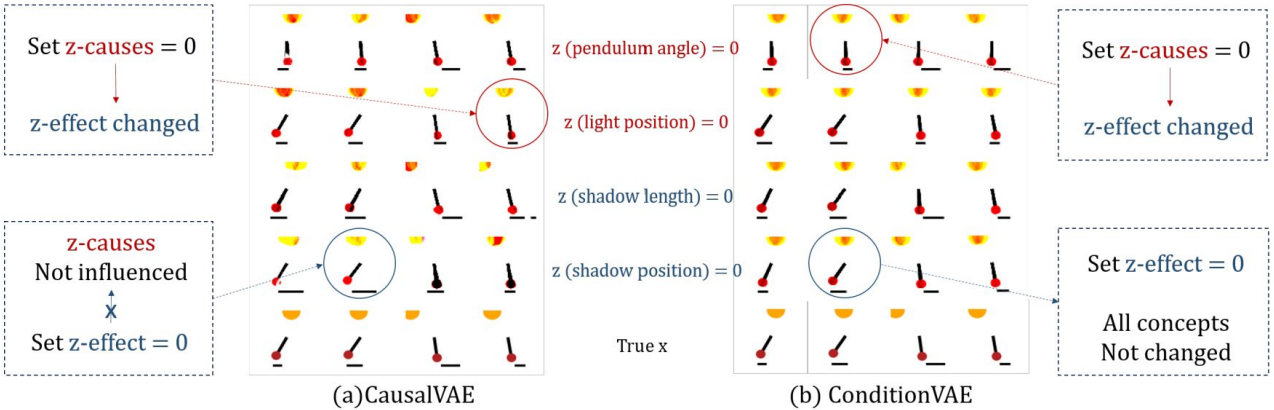
Causal disentanglement

- Learning disentanglement representation which align to real word concepts
- Learning causal graph automatically
- Achieve do-operation on casual representation
- Generate counterfactual images that do not appear in training data

The result of intervention

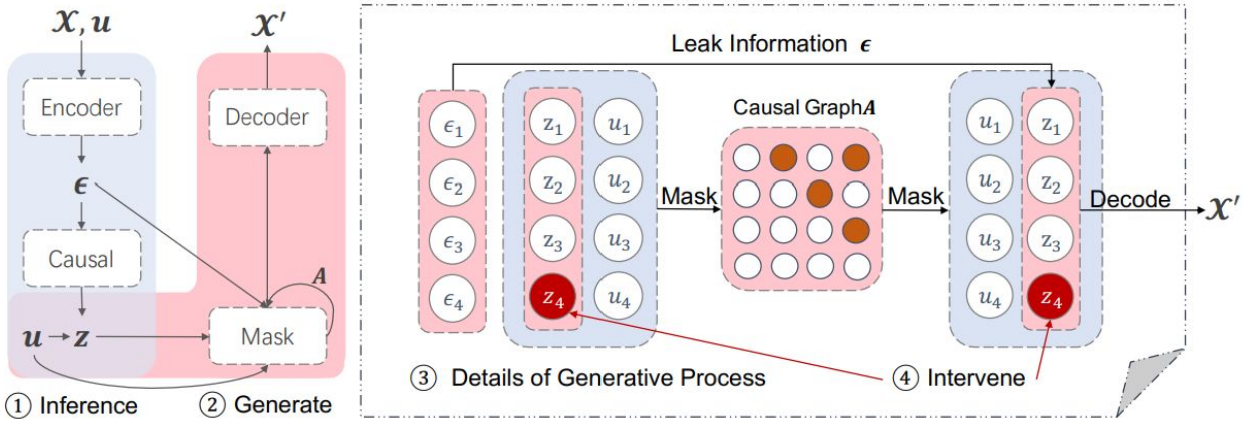


Causal graph



The result of intervention on pendulum dataset

Model structure of CausalVAE



$$\mathbf{z} = \mathbf{A}^T\mathbf{z} + \boldsymbol{\epsilon} = (I - \mathbf{A}^T)^{-1}\boldsymbol{\epsilon}, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

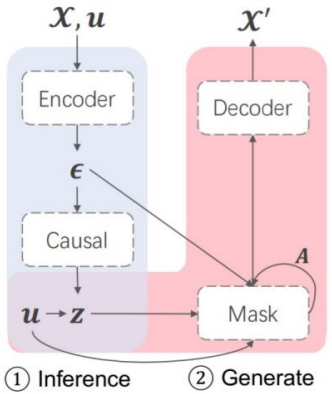
$$z_i = g_i(\mathbf{A}_i \circ \mathbf{z}; \boldsymbol{\eta}_i) + \epsilon_i,$$

Transforming Independent Exogenous Factors into Causal Representations

- Structural Causal Models (SCMs):

$$z = A^T z + \epsilon = \left(I - A^T\right)^{-1} \epsilon$$

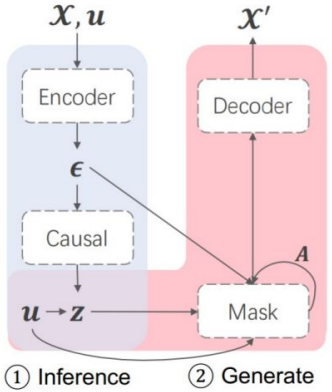
  - $\epsilon_i's$ are jointly independent which satisfy $\epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$.
  - z_i is the lower dimensional representation of i-th concept.
  - $\boldsymbol{A}$ is causal graph



VAE-based Model

- Inference Model $q_{\phi}(\mathbf{z}, \boldsymbol{\epsilon}|\mathbf{x}, \mathbf{u})$
  - Encoder transform observations $\boldsymbol{X}$ into $\boldsymbol{\epsilon}$
  - A Causal layer generate causal representation
    - $\boldsymbol{z} = \left(\boldsymbol{I} - \boldsymbol{A}^{\boldsymbol{T}}\right)^{-1}\boldsymbol{\epsilon}$
  - Introduce additional observation $\boldsymbol{u}$:
    - $\boldsymbol{z}$ satisfy conditional Gaussian $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{\lambda_1}(\boldsymbol{u}), \boldsymbol{\lambda_2^2}(\boldsymbol{u}))$, where $\boldsymbol{u}$ is additional observation.
- Generative Model

  $p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}, \boldsymbol{\epsilon}|\mathbf{u}) = p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}, \boldsymbol{\epsilon}, \mathbf{u}) p_{\boldsymbol{\theta}}(\boldsymbol{\epsilon}, \mathbf{z}|\mathbf{u})$
  - Introduce a Mask Layer
    - $z_i = g_i(\boldsymbol{A}^T\boldsymbol{z}; \boldsymbol{\eta_i}) + \epsilon_i$, where $\boldsymbol{\eta_i}$ is parameter of g_i
    - Achieve do-operation



Learning Strategy

- Evidence Lower Bound (ELBO) of CausalVAE

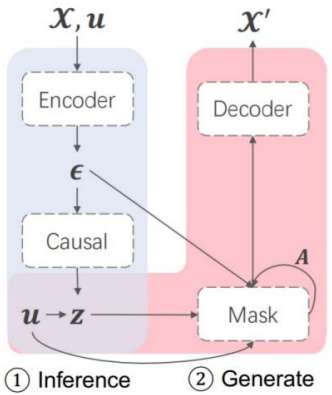
$$\mathbb{E}_{q_{\mathcal{X}}}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{u})] \geq \text{ELBO} = \mathbb{E}_{q_{\mathcal{X}}}[\mathbb{E}_{\boldsymbol{\epsilon},\mathbf{z}\sim q_{\boldsymbol{\phi}}}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z},\boldsymbol{\epsilon},\mathbf{u})] - \mathcal{D}(q_{\boldsymbol{\phi}}(\boldsymbol{\epsilon},\mathbf{z}|\mathbf{x},\mathbf{u})||p_{\boldsymbol{\theta}}(\boldsymbol{\epsilon},\mathbf{z}|\mathbf{u}))]$$

- Probabilistic definition:
  - Inference model： $q_{\boldsymbol{\phi}}(\boldsymbol{\epsilon},\mathbf{z}|\mathbf{x},\mathbf{u}) = q_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}|\mathbf{x},\mathbf{u})\delta(\boldsymbol{z} = \mathbf{C}\boldsymbol{\epsilon}) = q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x},\mathbf{u})\delta(\boldsymbol{\epsilon} = \mathbf{C}^{-1}\boldsymbol{z}),\ \mathbf{C} = (\mathbf{I} - \mathbf{A}^T)^{-1}$
  - Generative model: $p_{\boldsymbol{\theta}}(\boldsymbol{\epsilon},\mathbf{z}|\mathbf{u}) = p_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u}),$

- Decomposed ELBO

$$\text{ELBO} = \mathbb{E}_{q_{\mathcal{X}}}[\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x},\mathbf{u})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})] - \mathcal{D}(q_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}|\mathbf{x},\mathbf{u})||p_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})) - \mathcal{D}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x},\mathbf{u})||p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u}))].$$

$$p_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon}) = \mathcal{N}(\mathbf{0},\mathbf{I})$$
$$p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u}) = \Pi_i^n p_{\boldsymbol{\theta}}(z_i|u_i), p_{\boldsymbol{\theta}}(z_i|u_i) = \mathcal{N}(\lambda_1(u_i), \lambda_2^2(u_i)),$$



Loss function

$$l_u = \mathbb{E}_{q_{\mathcal{X}}} \|\mathbf{u} - \sigma(\mathbf{A}^T \mathbf{u})\|_2^2 \leq \kappa_1, \tag{11}$$

$$l_m = \mathbb{E}_{\mathbf{z} \sim q_{\boldsymbol{\phi}}} \sum_{i=1}^{n} \|z_i - g_i(\mathbf{A}_i \circ \mathbf{z}; \boldsymbol{\eta}_i)\|^2 \leq \kappa_2, \tag{12}$$

$$H(\mathbf{A}) \equiv tr((\mathbf{I} + \frac{c}{m}\mathbf{A} \circ \mathbf{A})^n) - n = 0, \tag{13}$$

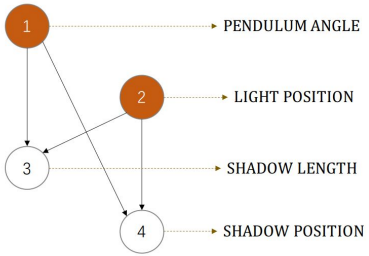
Constraints

maximize ELBO,
s.t. (11)(12)(13).

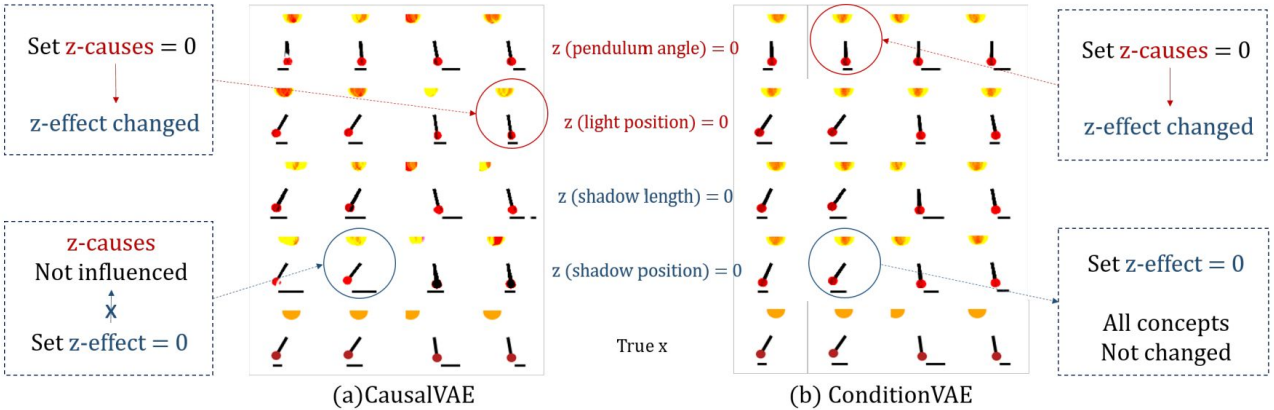
Langragian multiplier method

$$\mathcal{L} = -\text{ELBO} + \alpha H(\mathbf{A}) + \beta l_u + \gamma l_m,$$

The result of intervention

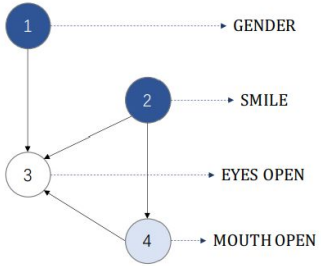


Causal graph

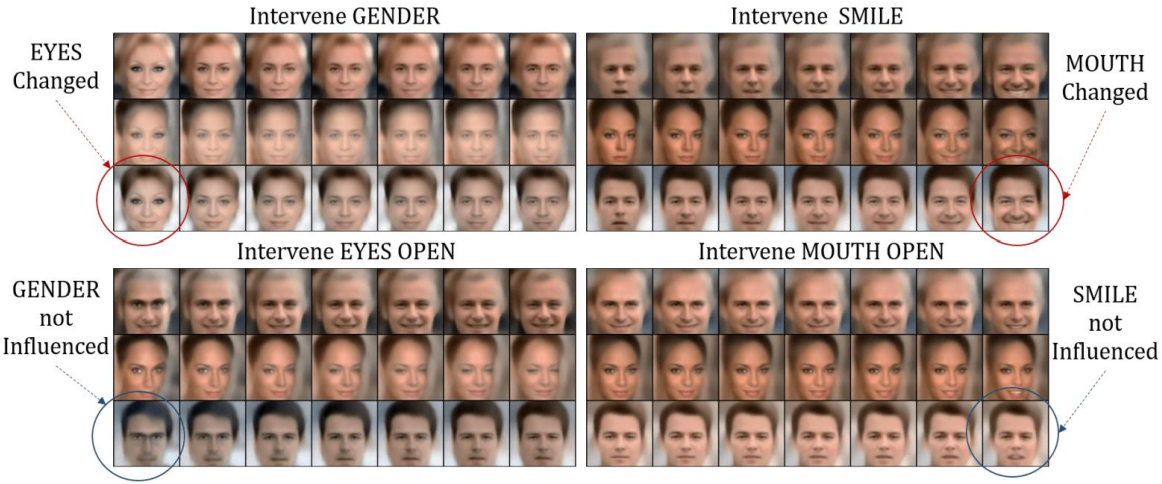


The result of intervention on pendulum dataset

Another example



Causal graph



Intervention results of CausalVAE model on CelebA (SMILE)

methods comparison based on MIC and TIC metrics

- Datasets (Synthetic : pendulum , flow
  Real world benchmark : CalebA)
- Supervised : conditionVAE
- Unsupervised : beta_vae , causalvae-unsup , ladderVae

The MIC and TIC between learned representation z and the label u

	CausalVAE		ConditionVAE		β-VAE		CausalVAE-unsup		LadderVAE	
Metrics(%)	MIC	TIC	MIC	TIC	MIC	TIC	MIC	TIC	MIC	TIC
Pendulum	**95.1 ±2.4**	**81.6 ±1.9**	93.8 ±3.3	80.5 ±1.4	22.6 ±4.6	12.5 ±2.2	21.2 ±1.4	12.0 ±1.0	22.4 ±3.1	12.8 ±1.2
Flow	72.1 ±1.3	56.4 ±1.6	**75.5 ±2.3**	**56.5 ±1.8**	23.6 ±3.2	12.5 ±0.6	22.8 ±2.7	12.4 ±1.4	34.3 ±4.3	24.4 ±1.5
CelebA(SMILE)	**83.7 ±6.2**	**71.6 ±7.2**	78.8 ±10.9	66.1 ±12.1	22.5 ±1.2	9.92 ±1.2	27.2 ±5.3	14.6 ±4.2	23.5 ±3.0	10.3 ±1.6
CelebA(BEARD)	**92.3 ±5.6**	**83.3 ±8.6**	89.8 ±6.2	78.7 ±7.7	22.4 ±1.9	9.82±2.2	11.4 ±1.5	20.0±2.2	23.5 ±3.0	8.1±1.2

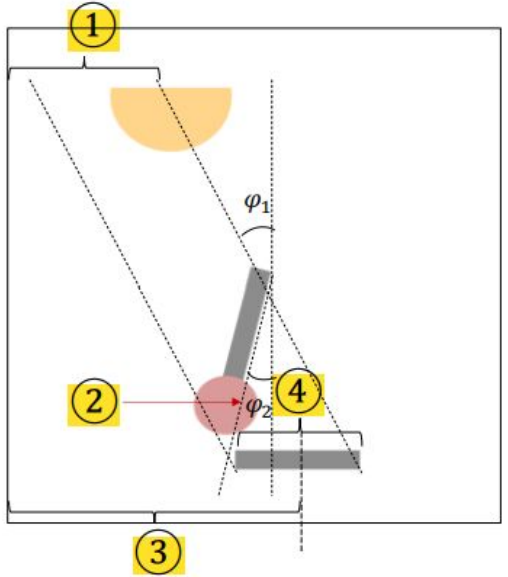
Conclusion

According to the authors, CausalVAE is the **first work on causal disentanglement**.

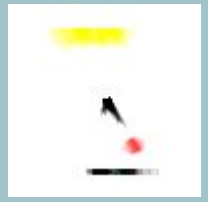
- It allows to discover causal relationships among the ground-truth factors
  - Prior knowledge can be eventually incorporated into the Adjacency matrix
- It is **identifiable**
- It supports the so called **do-operation**
- It considers **linear** causal relationships
- It requires **full knowledge of the ground-truth factors**

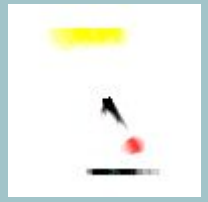
Code explanation

Data Preprocessing



output





Thanks for your attention