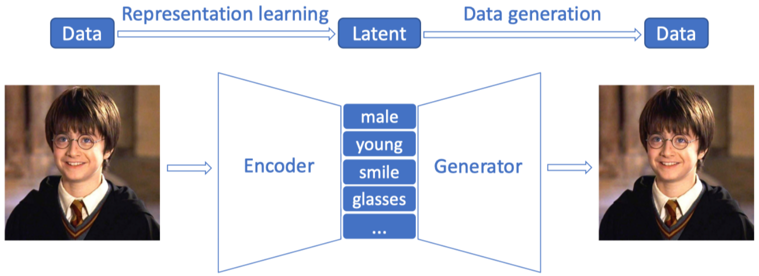
Disentangled Generative Causal Representation Learning

Xinwei Shen, Furui Liu, Hanze Dong, Qing Lian, Zhitang Chen, Tong Zhang

HKUST

Representation Learning and Disentanglement

Representation learning and generation



- Observed data $x \sim q_x$ on $\mathcal{X} \subseteq \mathbb{R}^d$
- Latent variable $z \sim p_z$ on $\mathcal{Z} \subseteq \mathbb{R}^k$
- Bidirectional generative model: learning an *encoder* $E : \mathcal{X} \to \mathcal{Z}$ (to learn representations) and a *generator* $G : \mathcal{Z} \to \mathcal{X}$ (to generate data).
- Example: variational auto-encoder (VAE)

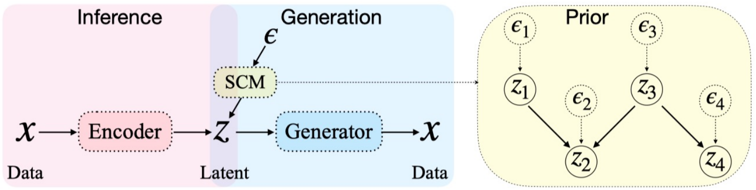
Disentanglement

Disentanglement as a common goal:

- In representation learning, an effective representation for downstream learning tasks should disentangle the underlying factors of variation.
- In generation, it is highly desirable if one can control the semantic generative factors.
- Both goals can be achieved with the *disentanglement* of latent variable z, which informally means that each dimension of z measures a distinct factor of variation in the data (Bengio et al., 2013).

Formulation

Generative model with a causal prior



- We adopt the general nonlinear Structural Causal Model (SCM):

$$f(z) = A^\top f(z) + h(\epsilon), \tag{3}$$

$$z = f^{-1}((I - A^\top)^{-1} h(\epsilon)) := F_\beta(\epsilon), \tag{4}$$

  where ϵ denotes the exogenous variables, $A \in \mathbb{R}^{k \times k}$ is the weighted adjacency matrix, f and h are element-wise nonlinear transformations.

- (3) enables intervention; (4) enables generation.

Supervised regularizer

- Let $\xi \in \mathbb{R}^m$ be the underlying factors of x, and y_i be some continuous or discrete observation of factor ξ_i satisfying $\xi_i = \mathbb{E}(y_i|x)$ for $i = 1, \dots, m$.
- Let $\bar{E}(x)$ be the deterministic part of the stochastic transformation $E(x)$, i.e., $\bar{E}(x) = \mathbb{E}(E(x)|x)$, which is used for representation learning.
- We consider the following objective:

$$L(E, G) = L_{\text{gen}}(E, G) + \lambda L_{\text{sup}}(E), \tag{2}$$

  where
  - $L_{\text{sup}} = \sum_{i=1}^m \mathbb{E}_{(x,y)}[\text{CE}(\bar{E}_i(x), y_i)]$ if y_i is the binary or bounded continuous label of ξ_i;
  - $L_{\text{sup}} = \sum_{i=1}^m \mathbb{E}_{(x,y)}[\bar{E}_i(x) - y_i]^2$ if y_i is the continuous observation of ξ_i.

Algorithm

Algorithm

Algorithm 1: Disentangled gEnerative cAusal Representation (DEAR) Learning

Input: training set $\{x_1, \ldots, x_N, y_1, \ldots, y_{N_s}\}$, initial parameter $\phi, \theta, \beta, \psi$, batch size n

while *not convergence* **do**

  for *multiple steps* **do**

    Sample $\{x_1, \ldots, x_n\}$ from the training set, $\{\epsilon_1, \ldots, \epsilon_n\}$ from $\mathcal{N}(0, I)$

    Generate from the causal prior $z_i = F_\beta(\epsilon_i), i = 1, \ldots n$

    Update ψ by descending the stochastic gradient:

    $\frac{1}{n}\sum_{i=1}^{n} \nabla_\psi \left[\log(1 + e^{-D_\psi(x_i, E_\phi(x_i))}) + \log(1 + e^{D_\psi(G_\theta(z_i), z_i)})\right]$

  Sample $\{x_1, \ldots, x_n, y_1, \ldots, y_{n_s}\}$, $\{\epsilon_1, \ldots, \epsilon_n\}$ as above; generate $z_i = F_\beta(\epsilon_i)$

  Compute θ-gradient: $-\frac{1}{n}\sum_{i=1}^{n} s(G_\theta(z_i), z_i)\nabla_\theta D_\psi(G_\theta(z_i), z_i)$

  Compute ϕ-gradient: $\frac{1}{n}\sum_{i=1}^{n} \nabla_\phi D_\psi(x_i, E_\phi(x_i)) + \frac{1}{n_s}\sum_{i=1}^{n_s} \nabla_\phi L_{\text{sup}}(\phi; x_i, y_i)$

  Compute β-gradient: $-\frac{1}{n}\sum_{i=1}^{n} s(G(z_i), z_i)\nabla_\beta D_\psi(G_\theta(F_\beta(\epsilon_i)), F_\beta(\epsilon_i))$

  Update parameters ϕ, θ, β using the gradients

Return: ϕ, θ, β

Formulation of DEAR

- Rewrite the generative loss:

$$L_{\text{gen}}(\phi, \theta, \beta) = D_{\text{KL}}(q_\phi(x, z), p_{\theta,\beta}(x, z)). \quad (5)$$

- Formulation to learn disentangled generative causal representations:

$$\min_{\phi,\theta,\beta} L(\phi, \theta, \beta) := L_{\text{gen}}(\phi, \theta, \beta) + \lambda L_{\text{sup}}(\phi). \quad (6)$$

Theory

Identifiability of disentanglement

Theorem

Assume the infinite capacity of E, G and f. Further assume the true binary adjacency matrix can be learned. Then DEAR learns the disentangled encoder E^*. Specifically, we have $g_i(\xi_i) = \sigma^{-1}(\xi_i)$ if CE loss is used in the supervised regularizer, and $g_i(\xi_i) = \xi_i$ if L_2 loss is used.

Optimization

- The SCM prior $p_\beta(z)$ and implicit generated conditional $p_\theta(x|z)$ make L_{gen} in (5) lose an analytic form.
- The lemma gives the gradient.
- We adopt a GAN method to adversarially estimate the gradient of L_{gen} as in Shen et al. (2020).

Lemma (Gradient)

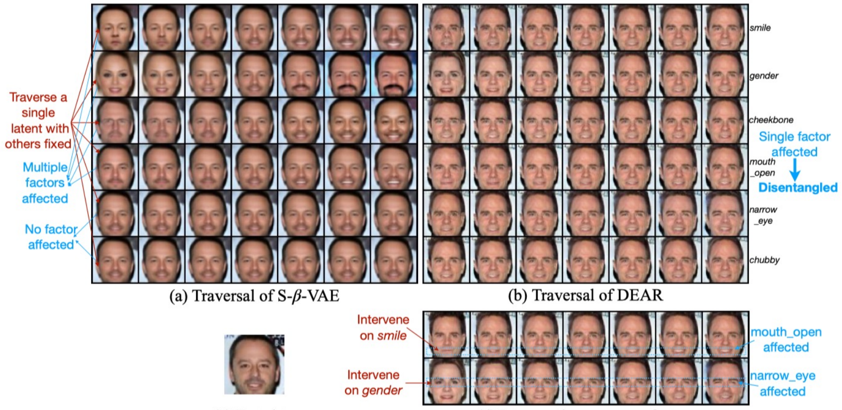
Let $r(x,z) = q(x,z)/p(x,z)$ *and* $\mathcal{D}(x,z) = \log r(x,z)$. *Then we have*

$$\begin{aligned}
\nabla_\theta L_{\text{gen}} &= -\mathbb{E}_{z\sim p_\beta(z)}[s(x,z)\nabla_x \mathcal{D}(x,z)^\top|_{x=G_\theta(z)}\nabla_\theta G_\theta(z)],\\
\nabla_\phi L_{\text{gen}} &= \mathbb{E}_{x\sim q_x}[\nabla_z \mathcal{D}(x,z)^\top|_{z=E_\phi(x)}\nabla_\phi E_\phi(x)],\\
\nabla_\beta L_{\text{gen}} &= -\mathbb{E}_\epsilon[s(x,z)(\nabla_x \mathcal{D}(x,z)^\top \nabla_\beta G(F_\beta(\epsilon)) + \nabla_z \mathcal{D}(x,z)^\top \nabla_\beta F_\beta(\epsilon))|_{z=F_\beta(\epsilon)}^{x=G(F_\beta(\epsilon))}],
\end{aligned} \tag{7}$$

where $s(x,z) = e^{\mathcal{D}(x,z)}$ *is the scaling factor.*

Controllable Generation

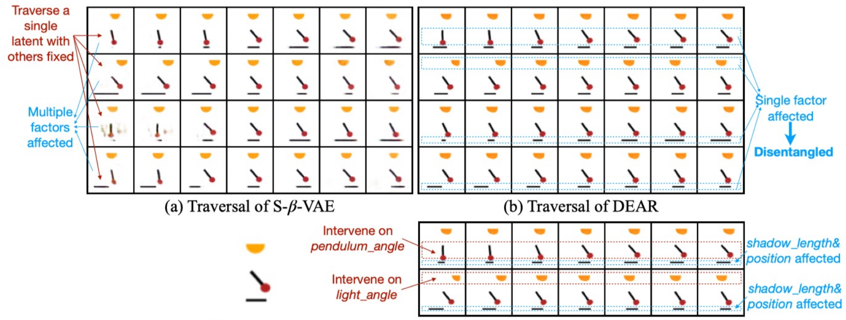
Causal controllable generation (CelebA)



(a) Traversal of S-β-VAE (b) Traversal of DEAR

(c) Test data (d) Intervention on cause factors

Causal controllable generation (Pendulum)



(a) Traversal of S-β-VAE (b) Traversal of DEAR

(c) Test data (d) Intervention on cause factors

Better Representations

Distributional robustness

Table: The worst-case and average test accuracy.

	(a) CelebA		(b) Pendulum	
Method	**WorstAcc(%)**	**AvgAcc(%)**	**WorstAcc(%)**	**AvgAcc(%)**
ERM	59.12±1.78	82.12±0.26	60.48±2.73	87.40±0.89
DEAR-lin-10%	71.40±0.47	81.04±0.14	63.93±1.33	89.70±0.63
DEAR-nlr-10%	70.44±1.02	81.94±0.31	65.59±1.90	90.19±0.63
ERM-multilabel	59.17±4.02	82.05±0.25	61.70±4.02	87.20±1.00
S-VAE	60.54±3.48	79.51±0.58	20.78±4.45	84.26±1.31
S-β-VAE	63.85±2.09	80.82±0.19	44.12±9.73	86.99±1.78
S-TCVAE	64.93±3.30	81.58±0.14	35.50±5.57	86.64±1.15
DEAR-lin	**76.05**±0.70	83.56±0.09	**74.95**±1.26	**93.61**±0.13
DEAR-nlr	71.37±0.66	**83.81**±0.08	72.48±0.74	93.11±0.14

Sample efficiency

- Statistical efficiency score: the average test accuracy based on 100 samples divided by the average accuracy based on 10,000/all samples (Locatello et al., 2019).

Table: Sample efficiency and test accuracy with different training sample sizes.

	(a) CelebA			(b) Pendulum		
Method	**100(%)**	**10,000(%)**	**Eff(%)**	**100(%)**	**all(%)**	**Eff(%)**
ResNet	68.06±0.19	79.51±0.31	85.59±0.27	79.71±0.98	90.64±1.57	87.97±2.11
DEAR-lin-10%	78.09±0.59	79.54±0.41	98.18±0.49	88.93±1.40	93.18±0.18	95.43±1.33
DEAR-nlr-10%	80.30±0.24	80.87±0.12	**99.29**±0.23	87.65±0.46	91.27±0.21	96.03±0.29
ResNet-pretrain	76.84±2.08	83.75±0.93	91.74±1.98	79.59±0.93	89.16±1.60	89.28±0.59
S-VAE	77.07±1.42	79.87±1.67	96.49±1.68	84.16±0.69	90.89±0.28	92.60±0.49
S-β-VAE	71.78±1.99	76.63±0.24	93.67±2.41	79.95±1.65	87.87±0.52	90.98±1.47
S-TCVAE	77.10±2.08	81.63±0.20	94.45±2.72	85.36±1.11	90.33±0.33	94.51±1.31
DEAR-lin	83.51±0.77	84.92±0.11	98.34±0.81	90.21±0.94	**93.31**±0.14	96.68±0.89
DEAR-nlr	**84.44**±0.48	**85.10**±0.09	99.23±0.51	**90.62**±0.32	92.57±0.08	**97.93**±0.29

Conclusion

- We identified a problem with previous methods using the independent prior assumption, and proved that they fail to disentangle when the underlying factors are causally correlated.
- We proposed a new disentangled learning method, DEAR, which integrates an SCM prior into a bidirectional generative model, trained with a suitable GAN loss.
- We provided theoretical justifications on the identifiability of the formulation and the asymptotic consistency of our algorithm.
- Extensive experiments were conducted to demonstrate the effectiveness of DEAR in causal controllable generation, and the benefits of the learned representations for downstream tasks.