Disentangled Generative Causal Representation Learning

Xinwei Shen

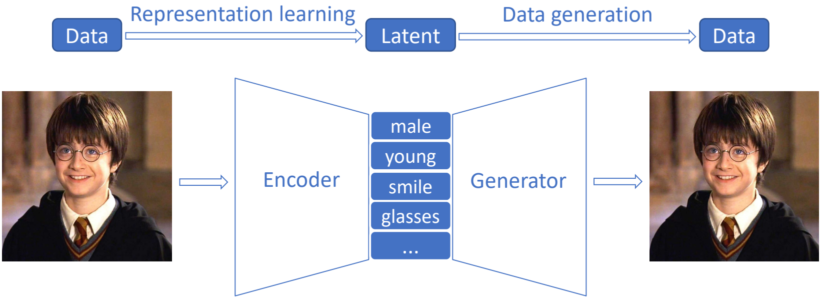
The Hong Kong University of Science and Technology

xinwei.shen@connect.ust.hk

March 9, 2021

Outline

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).

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).

How to achieve disentanglement?

Supervision is required

- Earlier unsupervised disentanglement methods mostly regularize the VAE objective to encourage independence of learned representations.

Supervision is required

- Earlier unsupervised disentanglement methods mostly regularize the VAE objective to encourage independence of learned representations.
- Locatello et al. (2019) show that unsupervised learning of disentangled representations is impossible: many existing unsupervised methods are actually brittle, requiring careful supervised hyperparameter tuning.

Supervision is required

- Earlier unsupervised disentanglement methods mostly regularize the VAE objective to encourage independence of learned representations.
- Locatello et al. (2019) show that unsupervised learning of disentangled representations is impossible: many existing unsupervised methods are actually brittle, requiring careful supervised hyperparameter tuning.
- To promote identifiability, recent work resorts to various forms of supervision.
- In this work, we also incorporate supervision on the ground-truth factors.

Causally correlated underlying factors

- Most existing methods are built on the assumption that the underlying factors are mutually independent.
- However, in many real world cases the semantically meaningful factors of interests are causally correlated, *i.e.*, connected by a causal graph.

Causally correlated underlying factors

- Most existing methods are built on the assumption that the underlying factors are mutually independent.
- However, in many real world cases the semantically meaningful factors of interests are causally correlated, *i.e.*, connected by a causal graph.
- We prove that methods with independent priors fail to disentangle causally correlated factors.
- Motivated by this finding, we propose a new method to learn Disentangled gEnerative cAusal Representations called DEAR.

Applications

- Causal controllable generation: to generate data from many desired interventional distributions of the latent factors.

Applications

- Causal controllable generation: to generate data from many desired interventional distributions of the latent factors.
- To use such representations in downstream tasks.
  - Disentangled: better sample complexity (Bengio et al., 2013).
  - Causal: invariant and thus robust under distribution shifts (Schölkopf, 2019).

Outline

Generative model

- Denote $(x, E(x)) \sim q_E(x, z), (G(z), z) \sim p_G(x, z)$.
- Consider the objective for generative modeling:

$$L_{\text{gen}}(E, G) = D_{\text{KL}}(q_E(x, z), p_G(x, z)), \tag{1}$$

which is equivalent to the VAE objective up to a constant.

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 $i = 1, \ldots, m$.
- 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.

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 $i = 1, \ldots, m$.
- 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.

Definition of a disentangled representation

- Intuitively, the above supervised regularizer aims at ensuring some alignment between factor ξ and latent variable z.

Definition (Disentangled representation)

Given the underlying factor $\xi \in \mathbb{R}^m$ of data x, a deterministic encoder E is said to learn a disentangled representation with respect to ξ if $\forall i = 1, \ldots, m$, there exists a 1-1 function g_i such that $E_i(x) = g_i(\xi_i)$. Further, a stochastic encoder E is said to be disentangled wrt ξ if its deterministic part $\bar{E}(x)$ is disentangled wrt ξ.

Unidentifiability with an independent prior

- Assumption: the underlying factors of interests are causally correlated, i.e., the elements of ξ are connected by a causal graph whose adjacency matrix A_0 is not a zero matrix.
- The following proposition indicates that the disentangled representation is generally unidentifiable with an independent prior.

Proposition

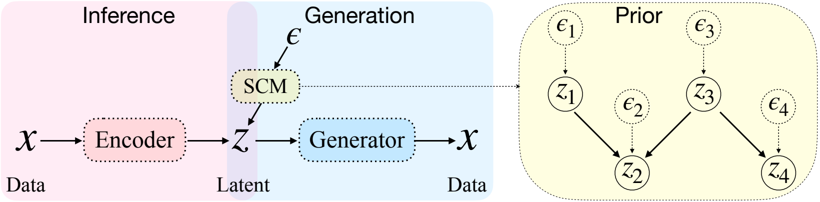
Let E^* be any encoder that is disentangled with respect to ξ. Let $b^* = L_{\text{sup}}(E^*)$, $a = \min_G L_{\text{gen}}(E^*, G)$, and $b = \min_{\{(E,G):L_{\text{gen}}=0\}} L_{\text{sup}}(E)$. Suppose the prior p_z is factorized, i.e., $p_z(z) = \prod_{i=1}^k p_i(z_i)$. Then we have $a > 0$, and either when $b^* \geq b$ or $b^* < b$ and $\lambda < \frac{a}{b-b^*}$, there exists a solution (E', G') so that E' is entangled and for any generator G, we have $L(E', G') < L(E^*, G)$.

Outline

Causal disentanglement learning

- Model
- Formulation
  - Theoretical justification (population)
- Optimization
- Algorithm
  - Theoretical justification (sample)

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.

Formulation of DEAR

- Rewrite the generative loss:

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

- Formulation to learn disentangled generative causal representations:

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

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.*

Algorithm

Algorithm 1: Disentangled gEnerative cAusal Representation (DEAR) Learning

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

while *not convergence* **do**

  for *multiple steps* **do**

    Sample $\{x_1, \dots, x_n\}$ from the training set, $\{\epsilon_1, \dots, \epsilon_n\}$ from $\mathcal{N}(0, I)$
    Generate from the causal prior $z_i = F_\beta(\epsilon_i), i = 1, \dots n$
    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 $\{x_1, \dots, x_n, y_1, \dots, y_{n_s}\}$, $\{\epsilon_1, \dots, \epsilon_n\}$ 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

Consistency of estimation

Theorem

Assume the objective function $L(\phi, \theta, \beta)$ in (6) is smooth and strongly convex, and achieves the global minimum at $(\phi^*, \theta^*, \beta^*)$. Under further appropriate conditions, there exists a sequence of $(N, N_s, N_d) \to \infty$ such that $(\hat{\phi}, \hat{\theta}, \hat{\beta}) \xrightarrow{p} (\phi^*, \theta^*, \beta^*)$.

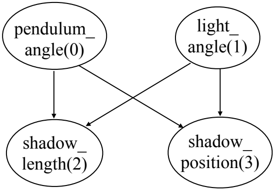
Outline

Datasets

Synthesized dataset Pendulum (Yang et al., 2020)

- Each image is generated by four continuous factors as shown in (b).
- We introduce 20% corrupted data whose shadow is randomly generated, mimicking some environmental disturbance.

(a)

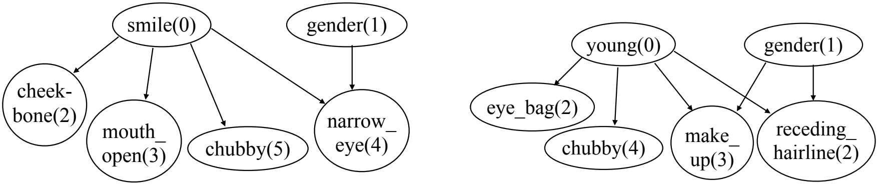


(b)

Datasets

CelebA (Liu et al., 2015)

- It contains 40 labelled binary attributes.
- We consider two groups of causally correlated factors.



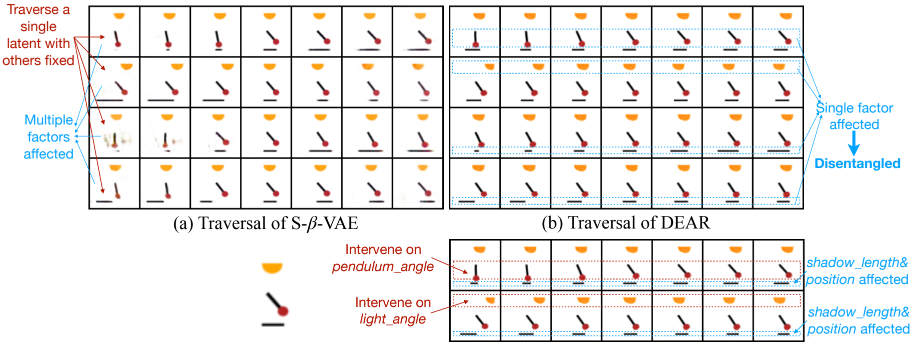
(a) CelebA-Smile (b) CelebA-Attractive

Figure: Underlying causal structures.

Causal controllable generation

- Traditional CG methods mainly manipulate the independent generative factors.
- With a learned SCM as the prior, we are able to generate images from many desired interventional distributions of the latent factors.

Causal controllable generation (Pendulum)



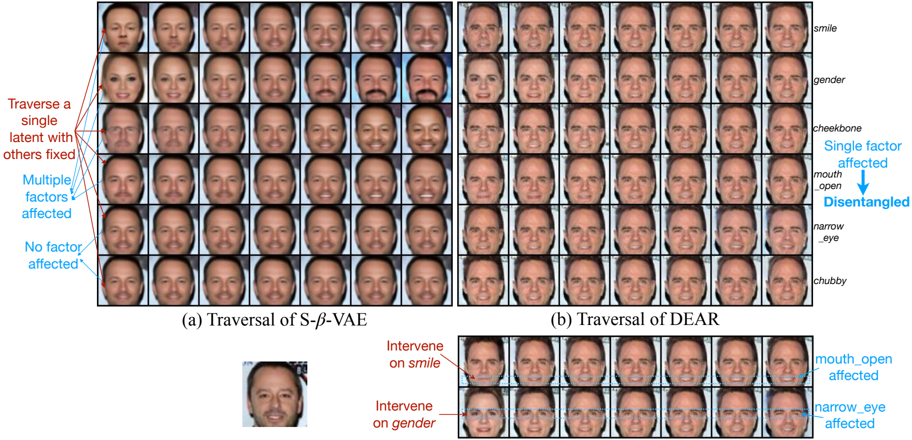
(a) Traversal of S-β-VAE

(b) Traversal of DEAR

(c) Test data

(d) Intervention on cause factors

Causal controllable generation (CelebA)



(a) Traversal of S-β-VAE

(b) Traversal of DEAR

(c) Test data

(d) Intervention on cause factors

Downstream task

- We consider some downstream prediction tasks.

Downstream task

- We consider some downstream prediction tasks.
- On CelebA, we consider the structure CelebA-Attractive. We artificially create a target label $\tau = 1$ if *young*=1, *gender*=0, *receding_hairline*=0, *make_up*=1, *chubby*=0, *eye_bag*=0, and $\tau = 0$ otherwise, indicating the attractiveness as a slim young woman with makeup and thick hair.

Downstream task

- We consider some downstream prediction tasks.
- On CelebA, we consider the structure CelebA-Attractive. We artificially create a target label $\tau = 1$ if *young*=1, *gender*=0, *receding_hairline*=0, *make_up*=1, *chubby*=0, *eye_bag*=0, and $\tau = 0$ otherwise, indicating the attractiveness as a slim young woman with makeup and thick hair.
- On the pendulum dataset, we regard the label of data corruption as the target τ, *i.e.*, $\tau = 1$ if the data is corrupted and $\tau = 0$ otherwise.

Downstream task

- We consider some downstream prediction tasks.
- On CelebA, we consider the structure CelebA-Attractive. We artificially create a target label $\tau = 1$ if *young*=1, *gender*=0, *receding_hairline*=0, *make_up*=1, *chubby*=0, *eye_bag*=0, and $\tau = 0$ otherwise, indicating the attractiveness as a slim young woman with makeup and thick hair.
- On the pendulum dataset, we regard the label of data corruption as the target τ, *i.e.*, $\tau = 1$ if the data is corrupted and $\tau = 0$ otherwise.
- In both cases, the factors to disentangle are causally related to τ, which are the features that humans use to do the task.
- A disentangled representation of these causal factors tends to be more data efficient and invariant to distribution shifts.

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

Distributional robustness

- We manipulate the training data such that the target label is more strongly correlated with the spurious attributes.

Distributional robustness

- We manipulate the training data such that the target label is more strongly correlated with the spurious attributes.
- On CelebA, we regard *mouth_open* as the spurious factor; on Pendulum, we choose *background_color* $\in \{\text{blue}(+), \text{white}(-)\}$.

Distributional robustness

- We manipulate the training data such that the target label is more strongly correlated with the spurious attributes.
- On CelebA, we regard *mouth_open* as the spurious factor; on Pendulum, we choose *background_color* $\in \{\text{blue}(+), \text{white}(-)\}$.
- Normal IID-based methods like ERM tend to exploit these easily learned spurious correlations in prediction.
- In contrast, causal factors are regarded invariant and thus robust under such shifts.

Distributional robustness

Table: The worst-case and average test accuracy.

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

Outline

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.

References

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, *35*(8), 1798–1828.

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of the ieee international conference on computer vision* (pp. 3730–3738).

Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., & Bachem, O. (2019, June). Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th international conference on machine learning (icml)* (Vol. 97, pp. 4114–4124). PMLR. Retrieved from http://proceedings.mlr.press/v97/locatello19a.html

Schölkopf, B. (2019). Causality for machine learning. *arXiv preprint arXiv:1911.10500*.

Shen, X., Zhang, T., & Chen, K. (2020). Bidirectional generative modeling using adversarial gradient estimation. *arXiv preprint arXiv:2002.09161*.

Yang, M., Liu, F., Chen, Z., Shen, X., Hao, J., & Wang, J. (2020). Causalvae: Structured causal disentanglement in variational autoencoder. *arXiv preprint arXiv:2004.08697*.

Thanks