CausalVAE: Structured Causal Disentanglement in Variational Autoencoder

Mengyue Yang [1] **Furui Liu** [2] **Zhitang Chen** [2] **Xinwei Shen** [3] **Jianye Hao** [2] **Jun Wang** [4]

Abstract

Learning disentanglement aims at finding a low dimensional representation, which consists of multiple explanatory and generative factors of the observational data. The framework of variational autoencoder is commonly used to disentangle independent factors from observations. However, in real scenarios, the factors with semantic meanings are not necessarily independent. Instead, there might be an underlying causal structure due to physics laws. We thus propose a new VAE based framework named CausalVAE, which includes causal layers to transform independent factors into causal factors that correspond to causally related concepts in data. We analyze the model identifiabitily of CausalVAE, showing that the generative model learned from the observational data recovers the true one up to a certain degree. Experiments are conducted on various datasets, including synthetic datasets consisting of pictures with multiple causally related objects abstracted from physical world, and a benchmark face dataset CelebA. The results show that the causal representations by CausalVAE are semantically interpretable, and lead to better results on downstream tasks. The new framework allows causal intervention, by which we can intervene any causal concepts to generate artificial data.

1. Introduction

Unsupervised disentangled representation learning is of importance in various applications such as speech, object recognition, natural language processing, recommender systems (Hsu et al., 2017; Ma et al., 2019; Hsieh et al., 2018). The reason is that it would help enhancing the performance of models, i.e. improving the generalizability, robustness against adversarial attacks as well as the explanability, by learning data's latent representation. One of the most common frameworks for disentangled representation learning is Variational Autoencoders (VAE), a deep generative model trained using backpropagation to disentangle the underlying explanatory factors. To achieve disentangling via VAE, one uses a penalty function to regularize the training of the model by reducing the gap between the distribution of the latent factors and a standard Multivatrate Gaussian. It is expected to recover the latent variables if the observations in real world are generated by countable independent factor. To further enhance the disentangement, a line of methods consider minimizing the mutual information between different latent factors. For example, Higgins et al. (2017); Burgess et al. (2018) adjust the hyperparameter to force latent codes to be independent of each other. Kim & Mnih (2018); Chen et al. (2018) further improve the independent by reducing total correlation.

The theory of disentangled representation learning is still at its early stage. We face problems such as the lack of a formal definition for disentangled representations and identifiability of disentanglement of generic models in unsupervised learning. To fill the gap, Higgins et al. (2018) proposed a new formalization of alignment between real world and latent space, and it is the first work which gives a formal definition of disentanglement. Locatello et al. (2018) challenged the common settings of state-of-the-arts, arguing that they can not find an identifiable model without inductive bias. Although they do consider the unreasonable aspect of disentanglement tasks, there are still unsolved problems like identifiability and explainability of the independent factors, or learnability of parameters from observations.

Common disentangling methods make a general assumption that the observations of real world are generated by countable independent factors. The recovered independent factors are considered good representations of data. We challenge this

[1]University of Chinese Academy of Sciences, Beijing, China [2]Noah's Ark Lab, Huawei, Shenzhen, China [3]The Hong Kong University of Science and Technology, Hong Kong, China [4]University College London, London, United Kingdom. Correspondence to: Furui Liu <liufurui2@huawei.com>.

assumption, as in many real world situations, meaningful factors are connected with causality.

Figure 1. A swinging pendulum.

Let us consider an example of a swinging pendulum Fig. 1, the direction of the light l and the pendulum p are causes of the location loc and length of shadow len. We aim at learning deep representations that correspond to the four concepts. Obviously, these concepts are not independent, i.e. the direction of the light and the pendulum determine the location and the length of the shadow. There exists various kinds of causal model which could measure this causal relationship i.e. Linear Structual Equation Models (SEM)(Shimizu et al., 2006). Existing methods for disentangled representation learning like β-VAE (Higgins et al., 2017) might not work as they forces the learned latent code to be as independent as possible. We argue the necessity to learn the causal representation as it allows us to intervention. For example, if we manage to learn latent codes corresponding to those four concepts, we can control the shape of the shadow without interrupting the generation of the light and the pendulum. This corresponds to the do-calculus (Pearl, 2009) in causality, where the system operates under the condition that certain variables are controlled by external forces.

In this paper, we develop a causal disentangled representation learning framework that recovers dependent factors by introduce Linear SEM into variation autoencoder framework. We enforce the structure to the learned latent code by designing a loss function that penalizes the deviation of the learned graph to a Directed Acyclic Graph (DAG). In addition, we analyze the identifiablilty of the proposed generative model, to guarantee an the learned disentangled codes are similar with the true one.

To verify the effectiveness of the proposed method, we conduct experiments on the dataset which consists of multiple causally related objects. We demonstrate empirically that The learned factors are with semantic meanings and can be intervened to generate artificial images that do not appear in training data.

We highlight our contributions of this paper as follows:

- We propose a new framework of generative model to achieve causal disentanglement learning.
- We develop a theory on identifiability of our generative models, which guarantees that the true generative model is recoverable up to certain degree.
- Experiments with synthetic and real world images are conducted to show the causal representations learned by proposed method have rich semantics and more effective for downstream tasks.

2. Related Works & Preliminary

In this section, we firstly provide background knowledge on disentangled representation learning, and we shall focus on recent state-of-the-arts using variational autoencoders. We review some recent advance of causality in generative models.

In the rest of the paper, we denote the latent variables by $\mathbf{z}$ with factorized density $p(\mathbf{z}) = \Pi_{i=1}^{d} p(z_i)$ where $d > 1$, and $p(\mathbf{z}|\mathbf{x})$ the posterior of the latent variables given the observation $\mathbf{x}$.

2.1. Disentanglement & Identifiability Problems

Disentanglement is a typical concept towards independent factorial representation of data. The classic method for identifying intrinsic independent factors is ICA (Comon, 1994; Jutten & Karhunen, 2003). Comon (1994) prove model identifiability of ICA in linear case. However, the identifiability of linear ICA model could not be extended to non-linear settings directly. Hyvarinen & Morioka (2016); Brakel & Bengio (2017) proposed a general identidfiability result for nonliear ICA, which links to the ideas of disentanglement under variational autoencoder. The disentangled representation learning learns mutually independent latent factors by an encoder-decoder framework. In the process, a standard normal distribution is used as prior

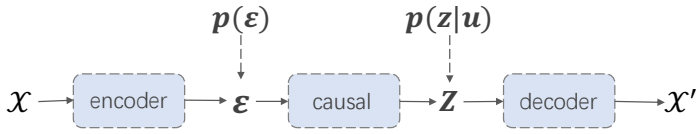


Figure 2. The information flow of CausalVAE. The observation $\mathbf{x}$ is the input of a encoder, and the encoder generates latent variable $\boldsymbol{\epsilon}$, whose prior distribution is assumed to be standard Multivariate Gaussian. Then it is transformed by the causal layer to be causal representation $\mathbf{z}$. The $\mathbf{z}$ are assumed to be with a conditional prior distribution $p(\mathbf{z}|\mathbf{u})$. $\mathbf{z}$ is taken as the input of the decoder to reconstruct observation $\mathbf{x}$.

of latent code. They use complex neural functions $q(z|x)$ to approximate parameterized conditional probability $p(z|x)$. This framework was extended by various existing works. Those works often introduce new independence constraints on the original loss function, leading to various disentangling metrics. β-VAE (Higgins et al., 2017) proposes an adaptation framework which adjusts the weight of KL term to balance between independence of disentangled factors and reconstruction performance. While factor VAE (Chen et al., 2018) proposes a new frame work which focuses solely on the independence of factors.

The aforementioned unsupervised algorithms do not perform well in some situations which content complex dependency among each factors, possibly because of lacking Inductive Bias and identidfiability of the generative model (Locatello et al., 2018).

The identidfiability problem in variational autoencoder are defined as follows: if the parameters $\tilde{\boldsymbol{\theta}}$ learned from data leads to a marginal distribution that equals the true one produced by $\boldsymbol{\theta}$, i. e., $p_{\tilde{\boldsymbol{\theta}}}(\mathbf{x}) = p_{\boldsymbol{\theta}}(\mathbf{x})$, then the joint distribution also matches $p_{\tilde{\boldsymbol{\theta}}}(\mathbf{x}, \mathbf{z}) = p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})$. It means that the learned parameters is identidfiability. Khemakhem et al. (2019) prove that the unsupervised variational autoencoder training results in infinite numbers of distinct models inducing the same data distributions, which means that the underlying ground truth is non-identifiable via unsupervised learning. On the contrary, by leveraging a few labels for supervision, one is able to recover the true model (Mathieu et al., 2018; Locatello et al., 2018). Kulkarni et al. (2015); Locatello et al. (2019) use few labels to guide model training to reduce the parameter uncertainty. Khemakhem et al. (2019) gives an identifiability result of variational autoencoder, by utilizing the theory of nonlinear ICA.

2.2. Causal Discovery from Pure Observational Data

We refer to causal representation as the representations that are structured by a causal graph. Discovering the causal graph from pure observational data has attracted large amount of attention in the past decades (Hoyer et al., 2009; Zhang & Hyvarinen, 2012; Shimizu et al., 2006). Pearl (2009) introduce a probabilistic graphical model based framework to learn causality from data. Shimizu et al. (2006) proposed an effective method called LiNGAM to learn the causal graph and they proved that the model is fully identifiable under the assumption that the causal relationship is linear and the noise is non-Gaussian distributed. Zheng et al. (2018) introduces DAG constraints for graph learning under continuous optimization (NOTEARS). Zhu & Chen (2019); Ng et al. (2019) use autoencoder framework to learn causal graph from data. Suter et al. (2018) use causality theories to explain disentangled latent representations. Furthermore, Zhang & Hyvarinen (2012) use more complex hypothesis function to represent a more sophisticated cause-effect relationships between two entities.

3. Method

In this section, we present our method by starting with a new definition of latent representation, and then give a framework of disentanglement using supervision. At last, we give theoretical analysis of the model identifiability.

3.1. Causal Model

To formalize causal representation framework, we consider n concepts in real world which have specific physical meanings. The concepts in observations are causally mixed by the causal relationship causal graph $\mathbf{A}$ *elaborate it in introduction.

As we mentioned, meaningful concepts are mostly not independent factors. We thus introduce causal representation in this paper. The causal representation is a latent data representation with a joint distribution that can be described by a

probabilistic graphical model, specially a Directed Acyclic Graph (DAG). We consider linear models in the paper, i.e. Linear Structural Equation models (SEM) on latent factors $\boldsymbol{z}$ as:

$$\mathbf{z} = \mathbf{A}^T\mathbf{z} + \boldsymbol{\epsilon} = (I - \mathbf{A}^T)^{-1}\boldsymbol{\epsilon}, \tag{1}$$

$$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{2}$$

where $\mathbf{z} \in \mathbb{R}^n$ is structural representation of n concepts. The independent noise are assumed to be Multivariate Gaussian. Once we are able to learn the causal representations from data, we are able to do intervention to the latent codes to generate artificial data which does not appear in the training data.

3.2. Generative Model

Our model is under the framework of VAE-based disentanglement. In additional to the encoder and the decoder structures, we introduce a causal layer to learn causal representations. The causal layer exactly implements a Linear SEM as described in Eq. 1, where $(\mathbf{I} - \mathbf{A}^T)^{-1}$ is the parameters to learn in this layer.

Unsupervised learning of the model might be infeasible due to the identifiability issue discussed in (Locatello et al., 2018). As a result, the learnability of the causal layer is in question, and predefined casual representation is not identifiable. To address this issue, similar to iVAE (Khemakhem et al., 2019), we use the additional information associated with the true causal concepts as supervising signals. The additional observations must include the information of real concepts like the label, pixel level observations. We build a causal conditional generative framework which uses the additional observations from causal concepts. We will discuss the identifiability of models given additional observations later.

We follow similar definition and notation to iVAE (Khemakhem et al., 2019). Denote by $\mathbf{x} \in \mathbb{R}^d$ the observed variables and $\mathbf{u} \in \mathbb{R}^n$ the additional information. u_i corresponds to the i-th concept in real causal system. Let $\mathbf{z} \in \mathbb{R}^n$ be the latent substantive variables with semantics and $\boldsymbol{\epsilon} \in \mathbb{R}^n$ be the latent independent variables where $\mathbf{z} = \mathbf{A}^T\mathbf{z} + \boldsymbol{\epsilon} = (\mathbf{I} - \mathbf{A}^T)^{-1}\boldsymbol{\epsilon}$. For simplicity, we denote $\mathbf{C} = (\mathbf{I} - \mathbf{A}^T)^{-1}$.

We now clarify the model assumptions for generation and inference process. Note that we regard both $\boldsymbol{z}$ and $\boldsymbol{\epsilon}$ as the latent variables. Consider the following conditional generative model parameterized by $\boldsymbol{\theta} = (\mathbf{f}, \mathbf{h}, \mathbf{C}, \mathbf{T}, \boldsymbol{\lambda})$:

$$p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}, \boldsymbol{\epsilon}|\mathbf{u}) = p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}, \boldsymbol{\epsilon}, \mathbf{u})p_{\boldsymbol{\theta}}(\boldsymbol{\epsilon}, \mathbf{z}|\mathbf{u}). \tag{3}$$

Let $\mathbf{f}(\mathbf{z})$ denotes the decoder which is assumed to be an invertible function and $\mathbf{h}(\mathbf{x})$ denote the encoder. Let $\boldsymbol{\epsilon} \in \mathbb{R}^n$ be independent noise variables, and $\mathbf{z} \in \mathbb{R}^n$ as the latent codes of n concepts.

We define the generation and inference process as follows:

$$\begin{aligned} p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}, \boldsymbol{\epsilon}, \mathbf{u}) = p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) = p_{\boldsymbol{\xi_{dec}}}(\mathbf{x} - \mathbf{f}(\mathbf{z})), \\ q_\phi(\epsilon|\mathbf{x}, \mathbf{u}) = p_{\boldsymbol{\xi_{enc}}}(\boldsymbol{\epsilon} - \mathbf{h}(\mathbf{x})). \end{aligned} \tag{4}$$

which is obtained by assuming the following decoding and encoding equations

$$\mathbf{x} = \mathbf{f}(\mathbf{z}) + \boldsymbol{\xi}_{dec}, \tag{5}$$

$$\boldsymbol{\epsilon} = \mathbf{h}(\mathbf{x}) + \boldsymbol{\xi}_{enc}, \tag{6}$$

where $\boldsymbol{\xi} = \{\boldsymbol{\xi}_{dec}, \boldsymbol{\xi}_{enc}\}$ are the vectors of independent noise with probability density $p_{\boldsymbol{\xi}}(\boldsymbol{\xi})$. When $\boldsymbol{\xi}$ is infinitesimal, the encoder and decoder distributions can be regarded as deterministic ones.

We define the joint prior $p_\theta(\boldsymbol{z}, \boldsymbol{\epsilon}|u)$ for latent variables $\boldsymbol{z}$ and $\boldsymbol{\epsilon}$ as

$$p_{\boldsymbol{\theta}}(\boldsymbol{\epsilon}, \mathbf{z}|\mathbf{u}) = p_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u}). \tag{7}$$

where $p_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$ and the prior of latent substantive variables $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u})$ is a factorized Gaussian distribution conditioning on the additional observation $\mathbf{u}$, i.e.

$$\begin{aligned} p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u}) = \Pi_i p_i(z_i|pa(z_i), u_i)p(u_i|pa(u_i)), \\ \sim \mathcal{N}(\mu_i(z_i)\lambda(u_i), \sigma_i(z_i)\lambda^2(u_i)). \end{aligned} \tag{8}$$

where λ is an arbitrary function (approximated by a neural network). In this paper, since each causal representation depend on the value of their parents node. We consider the case $\lambda(\mathbf{u}) = \mathbf{u}$ where $pa(u_i)$ denotes the parents node of u_i. The distribution has two sufficient statistics, the mean and variance of $\mathbf{z}$, which are denoted by $\mathbf{T}(\mathbf{z}) = (\boldsymbol{\mu}(\mathbf{z}), \boldsymbol{\sigma}(\mathbf{z})) = (T_{1,1}(z_1), \ldots, T_{n,2}(z_n))$.

3.3. Training Method

We apply variational Bayes to learn a tractable distribution $q_\phi(\boldsymbol{\epsilon}, \mathbf{z}|\mathbf{x}, \mathbf{u})$ to approximate the true posterior $p_{\boldsymbol{\theta}}(\boldsymbol{\epsilon}, \mathbf{z}|\mathbf{x}, \mathbf{u})$. Given data set D, we obtain empirical data distribution $q_D(\mathbf{x}, \mathbf{u})$. The parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ are learned by optimizing the following evidence lower bound (ELBO) on the expected data log-likelihood $\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{u}) = \int p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}, \boldsymbol{\epsilon}|\mathbf{u}) d\mathbf{z} d\boldsymbol{\epsilon}$:

$$\begin{aligned}\mathbb{E}_{q_D}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{u})] &\geq \text{ELBO} \\ &= \mathbb{E}_{q_D}[\mathbb{E}_{\boldsymbol{\epsilon}, \mathbf{u} \sim q_{\boldsymbol{\phi}}}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}, \boldsymbol{\epsilon}, \mathbf{u})] - D(q_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}, \mathbf{z}|\mathbf{x}, \mathbf{u})||p_{\boldsymbol{\theta}}(\boldsymbol{\epsilon}, \mathbf{z}|\mathbf{u}))].\end{aligned} \tag{9}$$

where $\mathcal{D}(\cdot\|\cdot)$ denotes KL divergence.

Noticing the one-to-one correspondence between $\boldsymbol{\epsilon}$ and $\mathbf{z}$, we simplify the variational posterior as follows:

$$\begin{aligned}q_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}, \mathbf{z}|\mathbf{x}, \mathbf{u}) &= q_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}|\mathbf{x}, \mathbf{u}) \mathbf{1}_{\mathbf{z}=\mathbf{C}\boldsymbol{\epsilon}}(\mathbf{z}), \\ &= q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}, \mathbf{u}) \mathbf{1}_{\boldsymbol{\epsilon}=\mathbf{C}^{-1}\mathbf{z}}(\boldsymbol{\epsilon}).\end{aligned}$$

Further according to the model assumptions introduced in Section 3.2, i.e., generation process (4) and prior (7), the ELBO can be rewritten as:

$$\text{ELBO} = \mathbb{E}_{q_D}[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{u})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})] - \mathcal{D}(q_\phi(\boldsymbol{\epsilon}|\mathbf{x}, \mathbf{u})||p_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})) - \mathcal{D}(q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{u})||p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u}))]. \tag{10}$$

where the third term is the key to disentangling the latent codes.

The causal adjacency matrix $\mathbf{A}$ is constrained to be a DAG. We introduce the acyclicity constraint. Instead of using traditional DAG constraint that is combinatorial, we adopt a continuous constraint function (Zheng et al., 2018; Zhu & Chen, 2019; Ng et al., 2019; Yu et al., 2019) . The function achieves 0 if and only if the adjacency matrix $\mathbf{A}$ are directed acyclic graph (Yu et al., 2019).

$$H(\mathbf{A}) \equiv tr((\mathbf{I} + \mathbf{A} \circ \mathbf{A})^n) - n = 0. \tag{11}$$

The decoder (generator) uses latent concept representation for reconstruction. To make learning process more smooth, we add the square term to the constraint. Thus the optimization of ELBO should be constrained by Eq. 11:

$$\begin{aligned}\text{maximize} \quad & \text{ELBO}. \\ \text{subject to} \quad & H(\mathbf{A}) = 0, \\ & H^2(\mathbf{A}) = 0.\end{aligned} \tag{12}$$

By lagrangian multiplier method, we have the new loss function

$$\mathcal{L} = -\text{ELBO} + \alpha(H(\mathbf{A}) + H^2(\mathbf{A})). \tag{13}$$

where α denotes regularization hyperparameters.

4. Identifiability Analysis

In this section, we present the identifiability of our proposed model. We adopt the $\sim$-*identifiability* (Khemakhem et al., 2019) as follows:

Definition 1. Let $\sim$ be the binary relation on Θ defined as follows:

$$\begin{aligned}(\mathbf{f}, \mathbf{h}, \mathbf{C}, \mathbf{T}, \boldsymbol{\lambda}) \sim (\tilde{\mathbf{f}}, \tilde{\mathbf{h}}, \tilde{\mathbf{C}}, \tilde{\mathbf{T}}, \tilde{\boldsymbol{\lambda}}) \Leftrightarrow \\ \exists \mathbf{B_1}, \mathbf{B_2} | \mathbf{T}(\mathbf{h}(\mathbf{x})) = \mathbf{B_1}\tilde{\mathbf{T}}(\tilde{\mathbf{h}}(\mathbf{x})), \mathbf{T}(\mathbf{f}^{-1}(\mathbf{x})) = \mathbf{B_2}\tilde{\mathbf{T}}(\tilde{\mathbf{f}}^{-1}(\mathbf{x})), \forall \mathbf{x} \in \mathcal{X}.\end{aligned} \tag{14}$$

If $\mathbf{B_1}$ is an invertible matrix and $\mathbf{B_2}$ is an invertible diagonal matrix in which each elements on diagonal correspond to u_i. we say that the model parameter is $\sim$-*identifiable*.

By extending Theorem 1 in iVAE (Khemakhem et al., 2019), we obtain the identifiability theory of our causal generative model.

Theorem 1. Assume that the data we observed are generated according Eq. 3-4 and the following assumptions hold,

1. The set $\{x \in \mathcal{X} | \phi_\xi(\mathbf{x}) = 0\}$ has measure zero, where ϕ_ξ is the characteristic function of the density p_ξ defined in Eq. 5.

2. The Jacobian matrix of decoder function $\mathbf{f}$ and encoder function $\mathbf{h}$ are full rank.

3. The sufficient statistics $T_{i,s}(z_i) \neq 0$ almost everywhere for all $1 \leq i \leq n$ and $1 \leq s \leq 2$, where $T_{i,s}(z_i)$ is the sth statistic of variable z_i.

4. The additional observations $u_i \neq 0$

. Then the parameters $(\mathbf{f}, \mathbf{h}, \mathbf{C}, \mathbf{T}, \boldsymbol{\lambda})$ are $\sim$-*identifiable*.

Sketch of proof:

Step 1: We analyze the identifiability of $\boldsymbol{\epsilon}$ started by $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{u}) = p_{\widetilde{\boldsymbol{\theta}}}(\mathbf{x}|\mathbf{u})$. Then we define a new invertible matrix $\mathbf{L}$ which contains additional observation u_i in causal system, and use it to prove that the learned $\tilde{\mathbf{T}}$ is the transformation of $\mathbf{T}$.

Step 2: We analyze the identifiability of $\mathbf{z}$ by replacing $\mathbf{C}\boldsymbol{\epsilon}$ in step 1 with $\mathbf{z}$. Then we use the invertible matrix $\mathbf{B_2}$, a diagonal matrix containing $\mathbf{u}$ to finish the proof.

More details are in **Appendix**.

The parameters $\boldsymbol{\theta}$ of true generative model are unknown during the learning process. The identifiablity of generative model is given by Theorem 1 which guarantees the parameters $\widetilde{\boldsymbol{\theta}}$ learned by hypothetical functions are in identifiable family.

In addition, all z_i in $\mathbf{z}$ align to the additional observation of concept i and they are expected to inherent the causal relationship of causal system. That is why that it could guarantee that the $\mathbf{z}$ are causal representations.

Then, for the causal representation $\mathbf{z}$ learned by the causal layer parameterized by $\mathbf{C}$, we here analyze the indentifiablity of $\mathbf{A}$.

Let $\mathbf{A}$ denote true causal structure of $\mathbf{z}$ and $\tilde{\mathbf{A}}$ denote the matrix leanred by our model. The following corollary illustrates the non-dentifiable $\mathbf{A}$.

Corollary 1. Suppose $\mathbf{A}$ and $\tilde{\mathbf{A}}$ are the true adjacency matrix and the adjacency matrix learned by our model, respectively. Then the following statement holds:

$$\mathbf{A} \sim \tilde{\mathbf{A}}. \tag{15}$$

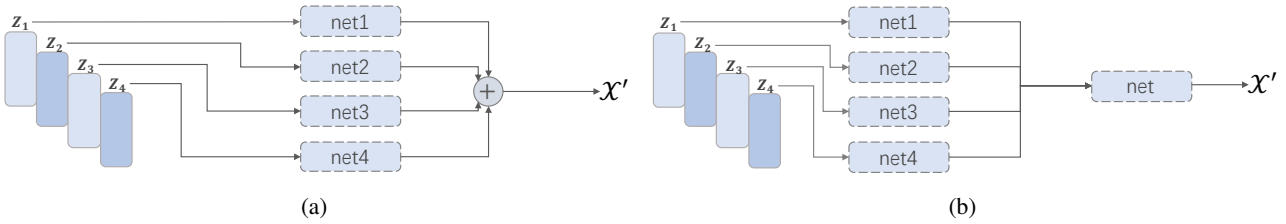
Or equivalently, the exists an invertible matrix $\mathbf{B}$ such that

$$\mathbf{T}(\mathbf{C}\mathbf{h}(\mathbf{x})) = \mathbf{B}\tilde{\mathbf{T}}(\tilde{\mathbf{C}}\tilde{\mathbf{h}}(\mathbf{x})). \tag{16}$$

Intuitively, the $\tilde{\mathbf{A}}$ learned in causal layer produces the $p(\mathbf{z}|\boldsymbol{\epsilon})$, which recovers the true one up to linear transformation.

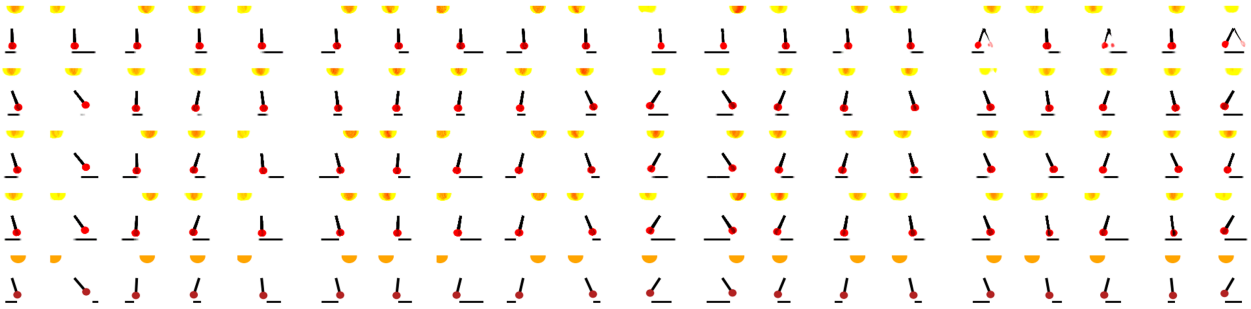
We furthur discuss some intuitions of idetifiability. Existing works often learn latent representation in an unsupervised way. However, our method uses the supervised ways, including additional observations. This supervision brings benefit that we can get the identifiability result of model.

The identifiability of the model under supervision of additional observation is obtained by the conditional prior $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u})$ generated from $\mathbf{u}$. The conditional prior guarantees that the sufficient statistics of $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u})$ are related to the value of $\mathbf{u}$. In other words, the values of $\mathbf{z}$ are determined by the supervision signal.



(a) (b)

Figure 3. Architectures of the two decoders used in experiments. (a) presents a structure that each concept is decoded separately by one network, and their results are assembled to be final output. (b) presents a structure that concepts are decoded by single neural network.



(a) CausalVAE-a (b) CausalVAE-b (c) DC-IGN-a (d) DC-IGN-b

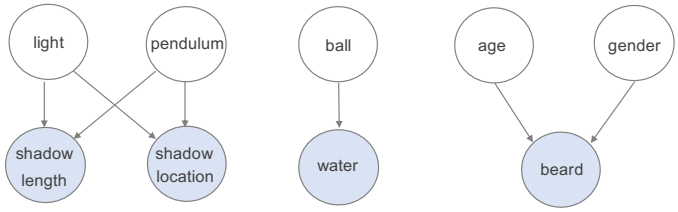
Figure 4. The results of DO-experiments on pendulum dataset. The first row presents the result of controlling the pendulum angles and the remaining rows are the results obtained by controlling light angle, shadow length, shadow location respectively. The bottom row is the true input image. Training epoch for models is set to be 100.

5. Experiments

In this section, we present the experimental results of our proposed method CausalVAE on datasets. Compared with those learned by the state-of-the-arts, the representation learned by our method performs well in both the synthetic causal image dataset and real world face data CelebA.

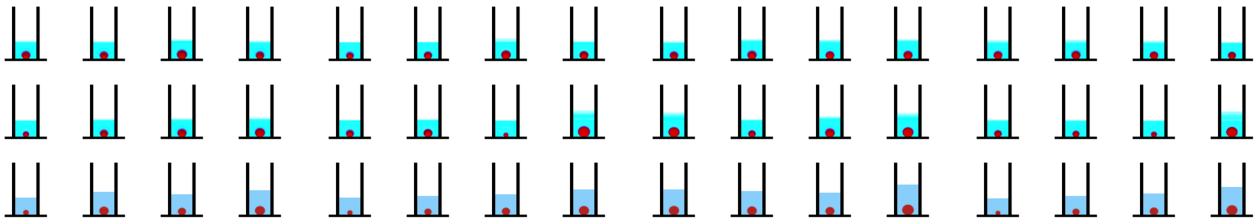
We test our CausalVAE on two tasks. The first task is factor interventions, and the second is downstream tasks, namely image classification.

In our experiments, the structure of the decoder largely influences the results. Thus, we use two designed decoders. The first one decodes the concepts separately and sum them up as the final output, and the second one decodes all concepts using a single neural network. The structures are in Fig. 3.



(a) (b) (c)

Figure 5. Causal graphs of three dataset. (a) shows the causal graph in pendulum dataset. The concepts are pendulum angle, light angle, shadow location and shadow length. (b) shows the causal graph in water dataset, on concepts water height and ball size. (c) shows the causal graph in CelebA, on concepts age, gender and beard.



(a) CausalVAE-a (b) CausalVAE-b (c) DC-IGN-a (d) DC-IGN-b

Figure 6. The results of DO-experiments on water. For each experiment we randomly choose 4 results. The first row presents results of controlling ball size (cause) and the second row controls water height (effect). The bottom one is the ground truth. Training epoch for models is set to be 100.

5.1. Dataset

5.1.1. Synthetic Data

We do experiments on the scenarios containing causally structured entities or concepts. We run models on a synthetic dataset, which include images consisting of causally related objects. A data generator is used to produce the images as model inputs. We will release our data generator soon.

Pendulum: We generate images with 3 entities (pendulum, light, shadow) which include 4 concepts (pendulum angle, light angle, shadow location, shadow length). The picture includes a pendulum. The angles of pendulum and the light are changing overtime. We use the projection laws to generate the shadows. The shadow are influenced by the light and angle of the pendulums. The causal graph of concepts is showed in Fig. 5 (a). In our experiments, we generate about 7k images (6K for training and 1k for inference), the angle of light and pendulum are ranged in around $[-\frac{\pi}{4}, -\frac{\pi}{4}]$.

Water: We produce artificial images, consisting of a ball in a cup filled with water. There are 2 concepts (ball size, height of water bar). The height is effect of the ball size. The causal graph is ploted in Fig. 5 (b) and the dataset includes 7k images, 6k images for training the disentanglemet model and the classifier model, and the rest of dataset are used as the test data of classifier.

5.1.2. Banchmark Dataset

In real world systems, cause and effect relationships commonly exist. To test our proposed method in these kinds of scenarios, we choose a banchmark CelebA[1], which is widely used in computer vision tasks. In this dataset, there are in total 200k human images with labels on different concepts. We focus on 3 concepts (age, gender and beard) on human faces in this dataset.

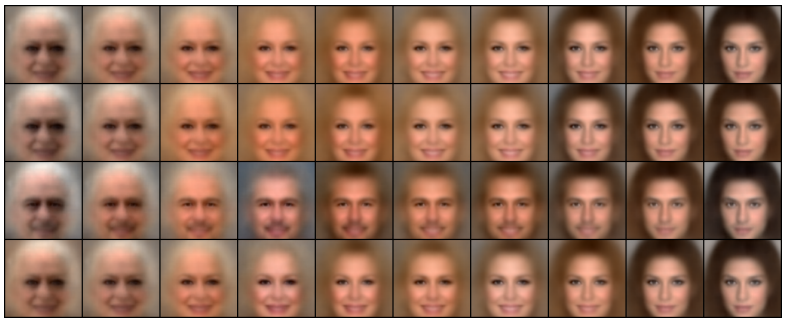
5.2. Baselines

CausalVAE-unsup: CausalVAE-unsup is the method under unsupervised setting. The architecture of the model is the same as CausalVAE but the additional observations are not used. We adjust the loss function by removing the additional observation.

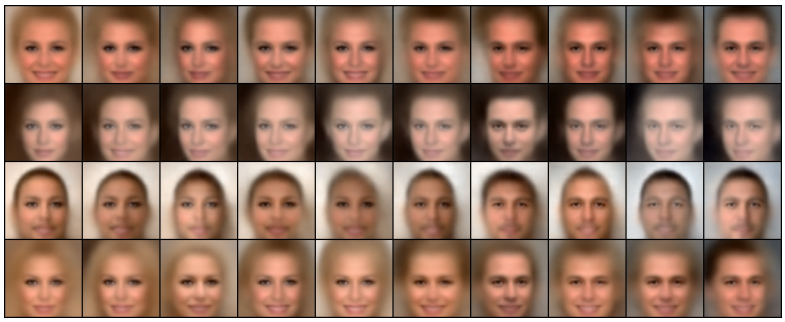
β-VAE: β-VAE is a common baseline for unsupervised disentanglement works. The dimensions of the latent representation are the same as that used in CausalVAE. The Standard Multivariate Gaussian distribution is adopted as the prior of latent variables.

DC-IGN: This baseline model is the model under supervised setting. They generate priors of latent variables conditional on the labels. As the case of β-VAE, dimensions of latent variables are set in line with our method.

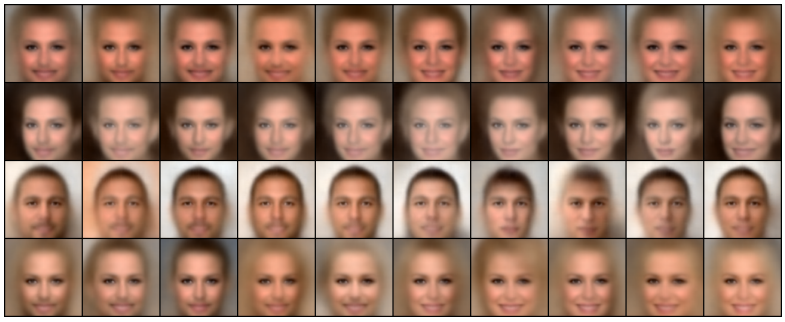
[1] http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html



(a) Age



(b) Gender



(c) Beard

Figure 7. Results of CausalVAE on CelebA, are results under hyperparameters $(\beta_1, \beta_2, \alpha) = (0.1, 0.2, 1)$. The controlled factors from top to bottom line are age, gender and beard, respectively. The first row shows the result of controlling gender, and the second row shows that of controlling age. The bottom is the result of controlling beard.

Model	Identifying cause labels				Identifying effect labels			
	pendulum decoder(a)	pendulum decoder(b)	water decoder(a)	water decoder(b)	pendulum decoder(a)	pendulum decoder(b)	water decoder(a)	water decoder(b)
β-VAE	0.6801	0.1905	0.7867	0.7707	0.6685	0.6679	0.7629	0.7629
DC-IGN	0.8313	0.7634	0.8570	0.8662	0.7649	0.8626	0.7710	0.7972
CausalVAE-unsup	0.8039	0.8028	**0.8667**	0.8496	**0.9362**	0.6663	0.7924	0.7990
CausalVAE	**0.8658**	**0.8587**	0.8564	**0.8656**	0.8952	**0.8874**	**0.8032**	**0.8038**

Table 1. The accuracy of classifiers on test dataset. The training epoach is 300 for pendulum and 50 for water. Experiments are repeated 5 times, and the median are reported.

5.3. Intervention experiments

Intervention experiments aim at testing if certain dimension of the latent codes has understandable semantic meanings. We control the value of latent vector by do-calculus operation introduced before, and check the reconstructed images.

For the experiments, all images of the dataset are used to train our proposed model CausalVAE and other baselines.

5.3.1. SYNTHETIC

For the experiments on synthetic dataset, we use different latent variable dimensions. We use 4 and 2 concepts on pendulum and water dataset, respectively. Then in all the experiments, we set the hyperparameter $\alpha = 1$.

We use CausalVAE-a to represent the CausalVAE model with decoder (a), and CausalVAE-b to represent the CausalVAE model with decoder (b). The same rules apply to the DC-IGN model.

We intervened 4 concepts of pendulum, and the results are showed in Fig. 4. The intervention strategy are illustrated in following step: 1) we learned a CausalVAE model; 2) we put a pendulum image into encoder and get the latent code $\mathbf{z}$. 3) we change the value of z_i as 0. For example, when we want to intervene gender, we will change the value of $z_{i=light}$ directly as 0 and keep other $z_{i\neq light}$ unchanged. 4) we put the total changed latent code $\mathbf{z}$ into decoder and got reconstruct image.

In implementation of CausalVAE, similar to β-VAE, we adjust the KL term in ELBO by multiplying a beta:

$$\beta_1 \mathcal{D}(q_\phi(\boldsymbol{\epsilon}|\mathbf{x})||p(\boldsymbol{\epsilon})) + \beta_2 \mathcal{D}(q_\phi(\mathbf{z}|\mathbf{x},\mathbf{u})||p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u}))$$

The hyperparameters of CausalVAE $\beta_1 = 0.1$, $\beta_2 = 0.3$.

Since we set the latent value as constant 0, if we controlled concept successfully, the pattern of controlled concept in one image will be the same as other images in its line. For example, when we control pendulum angel in 4(a), the first line shows that the pendulum angle in each images are almost same. And the same with light angle, each lights in different images of the second line are in the middle of top of images. And other concepts in line 3 and line 4 show similar effect.

From the results of CausalVAE with decoder (a) showed by Fig. 4(a), we find that the when we control the angle of light and pendulum, the location and length of shadows change correspondingly. But controlling the shadow factors, the light and pendulum are not affected. This result does not appear when we use decoder (b).

For experiments using decoder (b), controlling the two causes (pendulum angle and light angle), the two effects (shadow length and shadow location) do not change the reconstructed images in an expected way. In addition, controlling the effects factors in the latent representation does not influence the reconstructed images. The reason is that the decoder (b) itself may be an physical model which reasons out the effect factors based on the cause factors. The information contained in effect factors is hence not useful.

Then we analyze the results of DC-IGN. The intervention results are showed in Fig. 4(c) (d). Results show that there exists a problem that the control of causes sometimes does not influence the effects. This is because they do not have a causal layer to model the factors so that the learned factors are not concepts we expect.

We also test CausalVAE on water dataset. This scenario has two concepts. The intervention on the ball size (cause) influences the water height (effect), but the intervention on the effect does not influence the causes. We also find that the results have some fluctuations. The control of the concepts is not as good as that in the pendulum experiments. It is possibly because two concepts are related by a bijective function (one-to-one mapping), and it brings difficulty for the model to understand casual relations between concepts.

In water experiments, we also find that the decoder (a) performs better than decoder (b). We do not use the unsupervised method in these experiments because it will not guarantee all the representations are aligned to the concepts well.

5.3.2. HUMAN FACE

We also executed the experiments on real world banchmark data CelebA. In this kind of scenarios, the causal system is often complex, which has heterogeneous causes and effects. It is hard to observe all the concepts in the causal systems. In this experiments, we focus on only 3 concepts (age, gender and beard). Other concepts will possibly be confounders in system. Decoder (a) is used in our experiments.

We conducted our intervention experiments by following step: 1) we learned a CausalVAE model; 2) we put a human picture into encoder and get the latent code $\mathbf{z}$. 3) we change the value of z_i from -0.5 to 0.5, in which each z_i are correspond to the concept respectively. For example, when we want to intervene gender, we will change the value of $z_{i=gender}$ directly from -0.5 to 0.5 and keep other $z_{i\neq gender}$ unchanged. 4) we put the total changed latent code $\mathbf{z}$ into decoder and got reconstruct picture.

Different with synthetic data, we did not change the value of latent code as constant 0 but set the value in a range of number. Thus the figures will show the concept changing clearly.

The Fig. 7 demonstrate the result of CausalVAE under the parameters $\beta_1 = 0.1$, $\beta_2 = 0.2$. And (a)(b)(c) show the intervention experiments on concepts of age, gender and beard respectively. The interventions perform well that when we intervened the cause concept gender, not only the appearance of gender but the beard changed. In contrast, when we intervened effect concept beard, the gender in figure Fig. 7(c) are not changed.

5.4. Downstream Task

We also use the representation to do the downstream task on synthetic data. In this paper, we conduct tasks of image classification. We use the latent causal representation as the input of the classifier, and do experiments on predictions of causes and effects.

The 80% of dataset are used as the training data and the remaining are for testing. The cause conceptual vectors learned by our model are the inputs of a classifier, to predict either the cause labels or effect labels. The cause labels on pendulum dataset are produced by equally partitioning the angles 0 to 90 degree into 6 classes. The effect labels are constructed by dividing the original additional observations associated with the concepts into 3 classes. In water dataset, classifications on cause label and effect label are all binary classifications. The results are showed in table 1. It shows that using the latent codes learned by CausalVAE and DC-IGN, in general, leads to better classification performance than using that learned by other baselines. Our proposed method achieves the best performance. The choice of decoder does not have significant influences on the results when our model is used. However, it has a clear influence on the results of unsupervised baseline models like CausalVAE-unsup.

6. Conclusion

In this paper, we propose a framework for latent representation learning. We argue that causal representation is good representation for machine learning tasks, and incorporate a causal layer to learn this representation under the framework of variational autoencoder. We give identifiability result of the model when additional observations are available for supervised learning. The method is tested on synthetic and real datasets, on both intervention experiments and downstream tasks. Our viewpoint is expected to bring new insights into the domain of representation learning.

References

Brakel, P. and Bengio, Y. Learning independent features with adversarial nets for non-linear ica. *arXiv preprint arXiv:1710.05050*, 2017.

Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., and Lerchner, A. Understanding disentangling in $beta$-vae. *arXiv preprint arXiv:1804.03599*, 2018.

Chen, T. Q., Li, X., Grosse, R. B., and Duvenaud, D. K. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, pp. 2610–2620, 2018.

Comon, P. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. *Iclr*, 2(5):6, 2017.

Higgins, I., Amos, D., Pfau, D., Racaniere, S., Matthey, L., Rezende, D., and Lerchner, A. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.

Hoyer, P. O., Janzing, D., Mooij, J. M., Peters, J., and Schölkopf, B. Nonlinear causal discovery with additive noise models. In *Advances in neural information processing systems*, pp. 689–696, 2009.

Hsieh, J.-T., Liu, B., Huang, D.-A., Fei-Fei, L. F., and Niebles, J. C. Learning to decompose and disentangle representations for video prediction. In *Advances in Neural Information Processing Systems*, pp. 517–526, 2018.

Hsu, W.-N., Zhang, Y., and Glass, J. Unsupervised learning of disentangled and interpretable representations from sequential data. In *Advances in neural information processing systems*, pp. 1878–1889, 2017.

Hyvarinen, A. and Morioka, H. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. In *Advances in Neural Information Processing Systems*, pp. 3765–3773, 2016.

Jutten, C. and Karhunen, J. Advances in nonlinear blind source separation. In *Proc. of the 4th Int. Symp. on Independent Component Analysis and Blind Signal Separation (ICA2003)*, pp. 245–256, 2003.

Khemakhem, I., Kingma, D. P., and Hyvärinen, A. Variational autoencoders and nonlinear ICA: A unifying framework. *CoRR*, abs/1907.04809, 2019. URL http://arxiv.org/abs/1907.04809.

Kim, H. and Mnih, A. Disentangling by factorising. *arXiv preprint arXiv:1802.05983*, 2018.

Kulkarni, T. D., Whitney, W. F., Kohli, P., and Tenenbaum, J. Deep convolutional inverse graphics network. In *Advances in neural information processing systems*, pp. 2539–2547, 2015.

Locatello, F., Bauer, S., Lucic, M., Rätsch, G., Gelly, S., Schölkopf, B., and Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations. *arXiv preprint arXiv:1811.12359*, 2018.

Locatello, F., Tschannen, M., Bauer, S., Rätsch, G., Schölkopf, B., and Bachem, O. Disentangling factors of variation using few labels. *arXiv preprint arXiv:1905.01258*, 2019.

Ma, J., Zhou, C., Cui, P., Yang, H., and Zhu, W. Learning disentangled representations for recommendation. In *Advances in Neural Information Processing Systems*, pp. 5712–5723, 2019.

Mathieu, E., Rainforth, T., Siddharth, N., and Teh, Y. W. Disentangling disentanglement in variational autoencoders. *arXiv preprint arXiv:1812.02833*, 2018.

Ng, I., Zhu, S., Chen, Z., and Fang, Z. A graph autoencoder approach to causal structure learning. *CoRR*, abs/1911.07420, 2019. URL http://arxiv.org/abs/1911.07420.

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

Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7(Oct):2003–2030, 2006.

Sorrenson, P., Rother, C., and Köthe, U. Disentanglement by nonlinear ica with general incompressible-flow networks (gin). *arXiv preprint arXiv:2001.04872*, 2020.

Suter, R., Miladinović, D., Schölkopf, B., and Bauer, S. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. *arXiv preprint arXiv:1811.00007*, 2018.

Yu, Y., Chen, J., Gao, T., and Yu, M. Dag-gnn: Dag structure learning with graph neural networks. *arXiv preprint arXiv:1904.10098*, 2019.

Zhang, K. and Hyvarinen, A. On the identifiability of the post-nonlinear causal model. *arXiv preprint arXiv:1205.2599*, 2012.

Zheng, X., Aragam, B., Ravikumar, P. K., and Xing, E. P. Dags with no tears: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, pp. 9472–9483, 2018.

Zhu, S. and Chen, Z. Causal discovery with reinforcement learning. *CoRR*, abs/1906.04477, 2019. URL http://arxiv.org/abs/1906.04477.

A. Proof of Theorem 1

Based on information flow of the model, we would analyze the identifiability of $\boldsymbol{\epsilon}$ and $\mathbf{z}$. The general logic of the proofing follows (Khemakhem et al., 2019).

Step 1: Identifiability of $\boldsymbol{\epsilon}$.

Assume that $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{u})$ is equals to $p_{\widetilde{\boldsymbol{\theta}}}(\mathbf{x}|\mathbf{u})$. For all the observational pairs $(\mathbf{x}, \mathbf{u})$, let J_h denote the Jacobian matrix of the encoder function. There exist following equations,

$$\begin{aligned} p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{u}) &= p_{\widetilde{\boldsymbol{\theta}}}(\mathbf{x}|\mathbf{u}), \\ \Rightarrow \int_{\mathbf{z}} p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u}) d\mathbf{z} &= \int_{\mathbf{z}} p_{\widetilde{\boldsymbol{\theta}}}(\mathbf{x}|\mathbf{z}) p_{\widetilde{\boldsymbol{\theta}}}(\mathbf{z}|\mathbf{u}) d\mathbf{z}, \\ \Rightarrow \int_{\mathbf{x}'} p_{\boldsymbol{\theta}}(\mathbf{x}) p_{\boldsymbol{\theta}}(\mathbf{C}\mathbf{h}(\mathbf{x}')|\mathbf{u})|\det|\mathbf{C}|||\det(J_h(\mathbf{x}'))| d\mathbf{x}' &= \int_{\mathbf{x}'} p_{\boldsymbol{\theta}}(\mathbf{x}|\widetilde{\mathbf{C}}\widetilde{\mathbf{h}}(\mathbf{x}')) p_{\widetilde{\boldsymbol{\theta}}}(\widetilde{\mathbf{C}}\widetilde{\mathbf{h}}(\mathbf{x}')|\mathbf{u})|\det(\widetilde{\mathbf{C}})||\det(J_{\widetilde{\mathbf{h}}}(\mathbf{x}'))| d\mathbf{x}'. \end{aligned} \tag{17}$$

where $\mathbf{C} = (\mathbf{I} - \mathbf{A}^T)^{-1}$. In determining function $\mathbf{f}$ and $\mathbf{h}$, there exist a Gaussian distribution $p_{\boldsymbol{\xi}}(\boldsymbol{\xi})$ which has infinitesimal variance. Then, the $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{C}\mathbf{h}(\mathbf{x}'))$ can be written as $p_{\boldsymbol{\xi}}(\mathbf{x} - \mathbf{x}')$. As the assumption (1) holds, this term is vanished. Then in our method, there exists the following equation:

$$\begin{aligned} p_{\boldsymbol{\theta}}(\mathbf{C}\mathbf{h}(\mathbf{x}')|\mathbf{u})|\det|\mathbf{C}|||\det(J_{\mathbf{h}}(\mathbf{x}'))| &= p_{\widetilde{\boldsymbol{\theta}}}(\widetilde{\mathbf{C}}\widetilde{\mathbf{h}}(\mathbf{x}')|\mathbf{u})|\det(\widetilde{\mathbf{C}})||\det(J_{\widetilde{\mathbf{h}}}(\mathbf{x}'))|, \\ \Rightarrow \widetilde{p}_{\boldsymbol{\theta}}(\mathbf{x}) &= \widetilde{p}_{\widetilde{\boldsymbol{\theta}}}(\mathbf{x}). \end{aligned} \tag{18}$$

In Gaussian distribution, $p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u})$ can be written as follow:

$$p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{u}) = \Pi_i p_{\boldsymbol{\theta}}(z_i|pa(u_i), u_i) = \Pi_i p_{\boldsymbol{\theta}}(z_i|u_i). \tag{19}$$

where i is the concept index.

Adopting the definition of multivariate Gaussian distribution, we define

$$\boldsymbol{\lambda}_s(\mathbf{u}) = \begin{bmatrix} \lambda_1^s(u_1) & & \\ & \ddots & \\ & & \lambda_n^s(u_n) \end{bmatrix}. \tag{20}$$

There exists the following equations:

$$\begin{aligned} &\log|\det(\mathbf{C})| + \log|\det(J_{\mathbf{h}}(\mathbf{x}))| - \log \mathbf{Q}(\mathbf{C}\mathbf{h}(\mathbf{x})) + \sum_{s=1}^{2} \mathbf{T}_s(\mathbf{C}\mathbf{h}(\mathbf{x}))\boldsymbol{\lambda}_s(\mathbf{u}), \\ &= \log|\det(\widetilde{\mathbf{C}})| + \log|\det(J_{\widetilde{\mathbf{h}}}(\mathbf{x}))| - \log \widetilde{\mathbf{Q}}(\widetilde{\mathbf{C}}\widetilde{\mathbf{h}}(\mathbf{x})) + \sum_{s=1}^{2} \widetilde{\mathbf{T}}_s(\widetilde{\mathbf{C}}\widetilde{\mathbf{h}}(\mathbf{x}))\widetilde{\boldsymbol{\lambda}}_s(\mathbf{u}). \end{aligned} \tag{21}$$

where $\mathbf{Q}$ denotes the base measure. In Gaussian distribution, it is $\boldsymbol{\sigma}(\mathbf{z})$.

In learning process, $\widetilde{\mathbf{A}}$ is restricted as DAG. Thus, the $\widetilde{\mathbf{C}}$ exists which is full rank matrix. The item which is not related to u in Eq. 21 are cancelled out (Sorrenson et al., 2020).

$$\sum_{s=1}^{2} \mathbf{T}_s(\mathbf{C}\mathbf{h}(\mathbf{x}))\boldsymbol{\lambda}_s(\mathbf{u})) = \sum_{s=1}^{2} \widetilde{\mathbf{T}}_s(\widetilde{\mathbf{C}}\widetilde{\mathbf{h}}(\mathbf{x}))\widetilde{\boldsymbol{\lambda}}_s(\mathbf{u})). \tag{22}$$

where s denote the index of sufficient statistics of Gaussian distributions, indexing the mean (1) and the variance (2).

By assuming that the additional observation u_i is different, it is guaranteed that coefficients of the observations for different concepts are distinct. Thus, there exists an invertible matrix corresponding to additional information $\mathbf{u}$:

$$\mathbf{L} = \begin{bmatrix} \boldsymbol{\lambda}_1(\mathbf{u}) & \\ & \boldsymbol{\lambda}_2(\mathbf{u}) \end{bmatrix}. \tag{23}$$

Since the assumption that $u_i \neq 0$ holds, $\mathbf{L}$ is $2n \times 2n$ invertible and full rank diagonal matrix. We have:

$$\mathbf{B_3 L T}(\mathbf{h}(\mathbf{x})) = \widetilde{\mathbf{L}}\widetilde{\mathbf{T}}(\widetilde{\mathbf{h}}(\mathbf{x})) \Rightarrow \mathbf{T}(\mathbf{h}(\mathbf{x})) = \mathbf{B_1}\widetilde{\mathbf{T}}(\widetilde{\mathbf{h}}(\mathbf{x})). \tag{24}$$

where $\mathbf{B_3}$ is invertible matrix which corresponds to $\mathbf{C}$ and $\mathbf{B_1} = \mathbf{L^{-1}B_3^{-1}}\widetilde{\mathbf{L}}$. The definition of $\widetilde{\mathbf{L}}$ on learning model migrates the definition of $\mathbf{L}$ on ground truth.

Then we adopt the definitions following (Khemakhem et al., 2019). According to the Lemma 3 in (Khemakhem et al., 2019), we are able to pick out a pair $(\epsilon_i, \epsilon_i^2)$ such that, $(\mathbf{T}_i'(z_i), \mathbf{T}_i'(z_i^2))$ are linearly independent. Then concat the two points into a vector, and denote the Jacobian matrix $\mathbf{Q} = [J_{\mathbf{T}}(\boldsymbol{\epsilon}), J_{\mathbf{T}}(\boldsymbol{\epsilon}^2)]$, and define $\tilde{\mathbf{Q}}$ on $\tilde{\mathbf{T}}(\tilde{\mathbf{h}}^{-1} \circ \mathbf{h}(\boldsymbol{\epsilon}))$ in the same manner. By differentiating Eq. 24, we get

$$\mathbf{Q} = \mathbf{B_1}\tilde{\mathbf{Q}}. \tag{25}$$

Since the assumptiom (2) that Jacobian of $\mathbf{h}$ is full rank holds, it can prove that both $\mathbf{Q}$ and $\tilde{\mathbf{Q}}$ are invertible matrix. Thus from Eq. 25, $\mathbf{B_1}$ is invertible matrix. The details are shown in (Khemakhem et al., 2019).

Step 2: Under the assumption in Theorem 1, replace the $\mathbf{Ch}(\mathbf{x})$ with $\mathbf{f}^{-1}(\mathbf{x})$ in Eq. 17, then

$$\sum_{s=1}^{2} \mathbf{T}_s(\mathbf{f}^{-1}(\mathbf{x}))\boldsymbol{\lambda}_s(\mathbf{u})) = \sum_{s=1}^{2} \widetilde{\mathbf{T}}_s(\widetilde{\mathbf{f}}^{-1}(\mathbf{x}))\widetilde{\boldsymbol{\lambda}}_s(\mathbf{u})). \tag{26}$$

Then use Eq. 23 to replace the λ matrix in Eq. 26, and we get:

$$\mathbf{L}(\mathbf{f}^{-1}(\mathbf{x})) = \widetilde{\mathbf{L}}\widetilde{\mathbf{h}}(\widetilde{\mathbf{f}}^{-1}(\mathbf{x})), \tag{27}$$

$$\mathbf{T}(\mathbf{f}^{-1}(\mathbf{x})) = \mathbf{B_2}\widetilde{\mathbf{h}}(\widetilde{\mathbf{f}}^{-1}(\mathbf{x})). \tag{28}$$

where

$$\mathbf{B_2} = \begin{bmatrix} u_1^{-1}\widetilde{\lambda}_1^1(u_1) & & \\ & \ddots & \\ & & u_n^{-2}\widetilde{\lambda}_n^2(u_n) \end{bmatrix}. \tag{29}$$

Using the same way as shown in Eq. 25, it can prove that $\mathbf{B_2}$ is invertible matrix.

Eq. 24 and Eq. 28 both hold. Combining the two results supports the identifiability result in CausalVAE.

B. Implementation Details

We use one NVIDIA Tesla P40 GPU as our train and inference device.

For the implementation of CausalVAE and other baselines, we extend $\mathbf{z}$ to matrix $\mathbf{z} \in \mathbb{R}^{n*k}$ where n is the number of concepts and k is the latent dimension of each $\mathbf{z}_i$. The corresponding prior or conditional prior distributions of CausalVAE and other baselines are also adjusted (this means that we extend the multivariate Gaussian to the matrix Gaussian).

The subdiemnsions k for each synthetic (pendulum, water) experiments are set to be 4, and 16 for CelebA experiments. The implementation of continuous DAG constraint $H(\mathbf{A})$ follows the code of (Yu et al., 2019) [2].

B.1. DO-Experiments

In DO-experiments, we train the model on synthetic data for 100 epochs, on CelebA for 500 epochs and use this model to generate latent code of representations.

[2]https://github.com/fishmoon1234/DAG-GNN

B.1.1. SYNTHETIC

We present the experiments of our proposed CausalVAE with two kinds of decoder, and experiments of other baselines with decoder (a). The hyperparameters are defined as:

1. CausalVAE : $\beta_1 = 0.1, \beta_2 = 0.3$.
2. CausalVAE-unsup : $\beta_1 = 0.4$.
3. DC-IGN : $\beta_1 = 0.4$.
4. β-VAE : $\beta_1 = 0.4$.

From the figure, we find that the reconstruct errors of models with decoder (a) are higher than those with decoder (b).

The details of the neural networks are shown in Table 2.

B.1.2. CELEBA

The reconstruction errors during the training are shown in Fig.**??** (c). We only present the experiments with decode (a). The hyperparameters are:

1. CausalVAE : $\beta_1 = 0.1, \beta_2 = 0.2$.
2. CausalVAE-unsup : $\beta_1 = 0.3$.
3. DC-IGN : $\beta_1 = 0.3$.
4. β-VAE : $\beta_1 = 0.3$.

The details of the neural networks are shown in Table 3.

We also present the DO-experiments of CausalVAE and DC-IGN. In the training of the models, we both use face labels (age, gender and beard).

From the figures, we find that interventions on the latent variables constructed by CausalVAE, in general, show a better performance than on those constructed by DC-IGN, especially on cause concepts. The intervention on age in Fig. **??** is a good example demonstrating the performance.

When CausalVAE controls the effect latent variables beard, it will not change other concepts on the reconstructed images. However, in DO-experiments under DC-IGN, other conceptual parts like gender will change even though we only intervene on the beard dimension. This fact shows certain entanglement of the concepts learned by DC-IGN, and these concepts do not follow a cause-effect relationship.

B.2. Downstream Task

Here we show the loss curves during the training. We use 85% of the synthetic data for training. We present the experiments on two synthetic data and each one includes the experiments of identifying cause labels and effect labels. In addition, CausalVAE achieves better accuracy than most of the baselines. It shows evidence that our proposed method learns conceptual representations.

The network designs of the classifiers are shown in Table 4.

encoder	decoder(a)	decoder(b)
4*96*96×900 fc. 1ELU	concepts×(4× 300 fc. 1ELU)	concepts× (4× 300 fc. 1ELU)
900×300 fc. 1ELU	concepts× (300×300 fc. 1ELU)	concepts×(300×300 fc. 1ELU)
300×2*concepts*k fc.	concepts×(300× 1024 fc. 1ELU)	concepts×(300× 1024 fc.)
-	concepts×(1024× 4*96*96 fc.)	concepts×(1024× 4*96*96 fc.)

Table 2. Network design of models trained on synthetic data.

encoder	decoder
-	concepts×(1×1 conv. 128 1LReLU(0.2), stride 1)
4×4 conv. 32 1LReLU (0.2), stride 2	concepts×(4×4 convtranspose. 64 1LReLU (0.2), stride 1)
4×4 conv. 64 1LReLU (0.2), stride 2	concepts×(4×4 convtranspose. 64 1LReLU (0.2), stride 2)
4×4 conv. 64 1LReLU(0.2), stride 2	concepts×(4×4 convtranspose. 32 1LReLU (0.2), stride 2)
4×4 conv. 64 1LReLU (0.2), stride 2	concepts×(4×4 convtranspose. 32 1LReLU (0.2), stride 2)
4×4 conv. 256 1LReLU (0.2), stride 2	concepts×(4×4 convtranspose. 32 1LReLU (0.2), stride 2)
1×1 conv. 3, stride 1	concepts×(4×4 convtranspose. 3 , stride 2)

Table 3. Network design of models trained on CelebA.

pendulum-cause	pendulum-effect	water-cause	water-effect
4×50 fc. 1ELU	2×(4×32 fc. 1ELU)	4×32 fc. 1ELU	4×32 fc. 1ELU
32×6 fc.	2×(32×32 fc. 1ELU)	32×2 fc.	32×2 fc.
-	2×(32×3 fc.)	-	-

Table 4. Network designs of models for downstream tasks.