On Causally Disentangled Representations

Abbavaram Gowtham Reddy, Benin Godfrey L, Vineeth N Balasubramanian

Indian Institute of Technology Hyderabad, India
cs19resch11002@iith.ac.in, benin.godfrey@cse.iith.ac.in, vineethnb@iith.ac.in

Abstract

Representation learners that disentangle factors of variation have already proven to be important in addressing various real world concerns such as fairness and interpretability. Initially consisting of unsupervised models with independence assumptions, more recently, weak supervision and correlated features have been explored, but without a causal view of the generative process. In contrast, we work under the regime of a causal generative process where generative factors are either independent or can be potentially confounded by a set of observed or unobserved confounders. We present an analysis of disentangled representations through the notion of disentangled causal process. We motivate the need for new metrics and datasets to study causal disentanglement and propose two evaluation metrics and a dataset. We show that our metrics capture the desiderata of disentangled causal process. Finally, we perform an empirical study on state of the art disentangled representation learners using our metrics and dataset to evaluate them from causal perspective.

Introduction

Humans implicitly tend to use causal reasoning while learning and explaining real-world concepts. Deep learning models, however, are considered to be *black-box* (Lipton 2018) and also *correlational*, thus, we cannot directly rely on their decisions in safety-critical domains such as medicine, defence, aerospace, etc. Consequently, there has been a surge in using the ideas of causality to improve the learning and explanation capabilities of deep learning models in recent years (O' Shaughnessy et al. 2020; Suter et al. 2019; Goyal et al. 2019a,b; Chattopadhyay et al. 2019; Janzing 2019; Zmigrod et al. 2019; Pitis, Creager, and Garg 2020; Zhu, Ng, and Chen 2020; Schölkopf et al. 2021). Deep learning models that learn the underlying causal structures in data not only avoid this problem of learning purely correlational input-output relationships, but also help in providing causal explanations. In this work, we choose disentangled representation learning as a tool to study the usefulness of applying causality in machine learning.

Disentangled representation learning (Bengio, Courville, and Vincent 2013; Schölkopf et al. 2021) aims to identify the underlying independent generative factors of variation given an observed data distribution, and is an important problem to address given its applications to generalization (Montero et al. 2021), data generation (Zhu et al. 2020), explainability (Gilpin et al. 2018), fairness (Creager et al. 2019), etc. The generative processes underlying observational data often contain complex interactions among generative factors. Treating such interactions as *independent causal mechanisms* (Peters, Janzing, and Schölkopf 2017) is essential to many real-world applications including the development of learning algorithms that learn transferable mechanisms from one domain to another (Schölkopf et al. 2021).

The study of disentanglement in unsupervised settings, with independence assumptions on the generative factors, has been the dominant topic of study for some time in recent literature (Higgins et al. 2017; Kumar, Sattigeri, and Balakrishnan 2017; Kim and Mnih 2018; Chen et al. 2018). Considering the limitations of unsupervised disentanglement (Locatello et al. 2019) and potentially unrealistic nature of the independence assumptions, a few semi-supervised and weakly supervised disentanglement methods have also been developed more recently (Locatello et al. 2020; Chen and Batmanghelich 2020; Dittadi et al. 2021; Träuble et al. 2021). None of the above mentioned methods, however, take a causal view on the underlying data generative process while studying disentanglement. We study disentanglement from a causal perspective in this work, grounding ourselves on the very little work along this direction (Suter et al. 2019). Since causal generative processes can be complex with arbitrary depth and width in their graphical representations, we restrict ourselves to two-level causal generative processes of the form shown in Figure 1 as these, by itself, can model many real-world settings with confounding (Pearl 2009), and have not been studied before either in the context of disentanglement or representation learning. We then also study how well-known latent variable models – e.g., β-VAE (Higgins et al. 2017) – perform disentanglement in the presence of confounders.

To this end, based on the definition of a *disentangled causal process* by (Suter et al. 2019), we look at three essential properties of causal disentanglement and propose evaluation metrics that are grounded on the principles of causality to study the level of causal disentanglement achieved by a generative latent variable model. The analysis in (Suter et al. 2019) focused on a metric for interventional robustness, and was studied w.r.t. the encoder of a latent variable model, which limits us to operating on only the interventional distribution of the encoder output. We instead extend the definition of *disentangled causal process* to both the encoder and generator of a latent variable model. Studying disentanglement from the generator's perspective allows us to study the *counterfactual distribution* of the generator output along with the *interventional distribution* of the encoder output, thus enabling us to propose newer evaluation metrics to study causally disentangled representations.

Going further, given the limitations in existing datasets for study of causally disentangled representations – especially their realism, natural confounding, and complexity – we introduce a new realistic image dataset, CANDLE, whose generation follows a two-level causal generative process with confounders, considering our focus in this work. We also perform empirical studies on popular latent variable models to understand their ability to causally disentangle the underlying generative process using our metrics, our dataset as well as on existing datasets in this regard. CANDLE dataset, code, and the Appendix of this work are made publicly available at https://causal-disentanglement.github.io/IITH-CANDLE/. We summarize our key contributions below:

- We undertake a study of causal perspectives to disentanglement, and go beyond existing work to capture the generative process of latent variable representation learning models, and thus study interventional and counterfactual goodness.
- We present two new evaluation metrics to study disentangled representation learning that are consequences of the properties of causally disentangled latent variable models.
- We introduce a new image-based dataset that includes known causal generative factors as well as confounders to help study and improve deep generative latent variable models from a causal perspective.
- We perform empirical studies on various well-known latent variable models in this regard, analyze their performance from a causal perspective and also show how a small degree of weak supervision can help improve causally disentangled representation learning.

Related Work

Capturing the Generative Process. Evidently, the underlying causal generative process has an impact on understanding the level of disentanglement achieved by a model. For e.g., if two generative factors are correlated or confounded by external factors, existing models find it difficult to disentangle the underlying generative factors (Träuble et al. 2021; Dittadi et al. 2021). Much of the existing disentanglement literature relies on the assumption that generative factors are independent of each other (Higgins et al. 2017; Kim and Mnih 2018; Chen et al. 2018), and do not consider a causal view to the generating process. Recently, (Suter et al. 2019) presented a causal view to the generative process but focused on studying interventional robustness. We build on this work to present the desiderata of latent variable models to achieve causal disentanglement.

Disentanglement in Representation Learning. Disentangled representation learning has been largely studied in unsupervised generative models in the last few years (Higgins et al. 2017; Kumar, Sattigeri, and Balakrishnan 2017; Kim and Mnih 2018; Chen et al. 2018). These methods essentially assume that the learned generative (or latent) factors are independent. Recently, semi-supervised and weakly supervised methods have been proposed (Locatello et al. 2020; Chen and Batmanghelich 2020; Dittadi et al. 2021; Träuble et al. 2021) to achieve better disentanglement between the latent variables. However, these methods do not consider or study the alignment of such a learned disentangled representation to the causal generative model. Models that consider causal relationships among input features and learn structural causal models (Pearl 2009) in latent space have been proposed of late (Yang et al. 2020; Kocaoglu et al. 2018); however, such efforts have been far and few between, and evaluating the extent of causal disentanglement has not been the objective of such methods.

Evaluation Metrics for Disentanglement. Existing work on learning disentangled representations using latent variable models have largely developed their own metrics to evaluate the extent of disentanglement, including the BetaVAE metric (Higgins et al. 2017), FactorVAE metric (Kim and Mnih 2018), Mutual Information Gap (Chen et al. 2018), Modularity (Ridgeway and Mozer 2018), DCI Disentanglement (Eastwood and Williams 2018), and the SAP Score (Kumar, Sattigeri, and Balakrishnan 2017). One important drawback of these metrics is that the possible effects of confounding in a generative process are not considered. Confounding is a critical aspect of real-world generative processes where the relationship between two variables can in turn depend on other variables (called *confounding variables* or *confounders*, see Figure 1), which could either be observed or unobserved. Confounders are the reasons to observe spurious correlations among generative factors in observational data. This is one of the primary challenges in studying causal effects, and requires careful consideration when evaluating disentangled representations. The first causal effort in this direction was the Interventional Robustness Score (IRS) developed by (Suter et al. 2019), which however relies exclusively on the learned latent space to evaluate disentangled representations. The IRS metric allows for presence of confounders in the data generating process, but does not make an effort to differentiate them in the learned latent variable space (e.g., two generative factors that are highly correlated can still be encoded by a single latent factor, which can be limiting). We empirically observe that one can get a good IRS score with very little training (please see Appendix) but at the cost of bad reconstructions, i.e. the IRS metric does not capture the goodness of the disentangled latent variables in generating useful data. Good reconstructions and thus good counterfactual generations are equally important in our quest to achieve deep learning models that learn causal generative factors. Our proposed evaluation metrics address this important issue by penalizing the latent variables that are confounded and by quantitatively evaluating the generated counterfactuals.

Image Datasets for Study of Disentanglement. Image

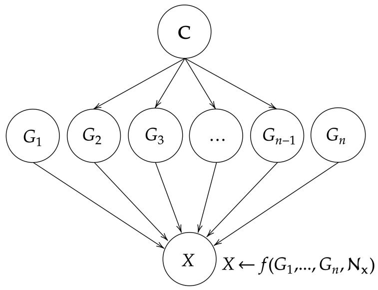

Figure 1: A causal process for generating X with generative factors $\{G_1, \ldots, G_n\}$ and confounders C

datasets that are studied in disentangled representation learning include dSprites (Matthey et al. 2017), smallNORB (LeCun, Huang, and Bottou 2004), 3Dshapes (Burgess and Kim 2018), cars3D (Fidler, Dickinson, and Urtasun 2012), MPI3D (Gondal et al. 2019), Falcor3D, and Isaac3D (Nie et al. 2020). These datasets, which are mostly synthetic, are generated based on a causal graph in which all factors of variation are assumed to be independent, and the causal graph is largely one-level. We introduce a realistic image dataset that involves two-level causal graphs with semantically relevant confounders to study various disentanglement methods using our proposed metrics. More details including comparisons with existing datasets are presented in Section and in the Appendix.

Disentangled Causal Process

We work under the regime of causal generative processes of the form shown in Figure 1 where a set of generative factors $\mathbf{G} = \{G_1, G_2, \ldots, G_n\}$ are independent by nature but can potentially be confounded by a set of confounders ($\mathbf{C}$). To this end, we begin by stating the definition of *disentangled causal process* (Suter et al. 2019) below.

Definition 1. (Disentangled Causal Process (Suter et al. 2019)). When a set of generative factors $\mathbf{G} = \{G_1, \ldots, G_n\}$ do not causally influence each other (i.e., $G_i \nrightarrow G_j$) but can be confounded by a set of confounders $\mathbf{C} = \{C_1, \ldots, C_l\}$, a causal model for X with generative factors $\mathbf{G}$ is said to be disentangled if and only if it can be described by a *structural causal model* (Pearl 2009) of the form:

$$\begin{aligned} C_j &\leftarrow \mathcal{N}_{c_j}; j \in \{1, \ldots, l\} \\ G_i &\leftarrow g_i(PA_i^C, \mathcal{N}_{G_i}); i \in \{1, \ldots, n\} \\ X &\leftarrow f(G_1, \ldots, G_n, \mathcal{N}_x) \end{aligned}$$

where f, g_i are independent causal mechanisms, $PA_i^C \subseteq \{C_1, \ldots, C_l\}$ represents the parents of G_i and $\mathcal{N}_{c_j}, \mathcal{N}_{G_i}, \mathcal{N}_x$ are independent noise variables.

We now examine an essential property (Property 1) of such a disentangled causal process and we extend it to latent variable models to be able to propose new evaluation metrics for causal disentanglement.

Property 1. In a disentangled causal process of type shown in Figure 1, G_i does not causally influence $G_j, i \neq j$ because any intervention on G_i will remove incoming edges from $\mathbf{C}$ and X is a collider in the path $G_i \rightarrow X \leftarrow G_j$ (Pearl 2009). As a consequence, G_i will not have any causal effect on X via G_j and all the causal effect of G_i on X is via the directed edge $G_i \rightarrow X$ (Suter et al. 2019).

We now translate Property 1 to deep generative latent variable models. Considering their well-known use in disentanglement literature, we focus on Variational Auto-Encoders (VAEs) for this purpose. A latent variable model $\mathcal{M}(e, g, p_X)$ with an encoder e, generator g and a data distribution p_X, assumes a prior $p(Z)$ on the latent space, and a generator g (often a deep neural network) is parametrized as $p_\theta(X|Z)$. We then approximate the posterior $p(Z|X)$ using a variational distribution $q_\phi(Z|X)$ parametrized by another deep neural network, called the encoder e. The prior is usually assumed to be an isotropic Gaussian (Kingma and Welling 2013; Rezende, Mohamed, and Wierstra 2014) and the model is trained on $x \sim p_X$ by maximizing the log-likelihood of reconstruction and minimizing the difference between the prior and approximated posterior. This leads to a set of generative factors $\mathbf{G}$ encoded as a set of latent dimensions $\mathbf{Z} = \{Z_1, \ldots, Z_m\}$. Specifically the latent variable model captures each generative factor G_i as a set of latent dimensions $\mathbf{Z}_I \subseteq \mathbf{Z}$ ($\mathbf{Z}$ indexed by a set of indices I). Ideally, one would want I to be a singleton set so each generative factor has a unique latent variable learned by the model, but it is also possible for such a model to encode G_i into more than one latent dimension (e.g., an angle can be encoded as $\sin\theta, \cos\theta$ in two different latent dimensions (Ridgeway and Mozer 2018)). In order for latent variable models to view the generator g as a causal mechanism to generate observations $\hat{x}$, $\mathbf{Z}$ acts now as a proxy for the true generative factors $\mathbf{G}$ (we use x for an instance of a random variable X, $\hat{x}$ hence denotes the reconstruction of x obtained through the generator g).

As a consequence of Property 1, any latent variable model $\mathcal{M}$ should satisfy the following two properties to achieve causal disentanglement.

Property 2. If a latent variable model $\mathcal{M}(e, g, p_X)$ disentangles a causal process of type shown in Figure 1, and the encoder e learns a latent space $\mathbf{Z}$ such that each generative factor G_i is mapped to a unique $\mathbf{Z}_I$ (unique $\mathbf{Z}_I$ refers to the scenario: $\mathbf{Z}_I \cap \mathbf{Z}_J = \emptyset, I \neq J, |I|, |J| \geq 0$ where $\mathbf{Z}_J$ is responsible for another generative factor G_j), then the generator g is a disentangled causal mechanism that models the underlying generative process.

Property 2 is similar to *encoder disentanglement* in (Shu et al. 2020) but we view it in terms of the generator than the encoder. Property 2 essentially boils down to learning a one-to-one mapping between each G_i and $\mathbf{Z}_I$, i.e. when two data points x_1, x_2 differ in only one generative factor G_i, one should observe a change only in $\mathbf{Z}_I$ when generating $\hat{x}_1, \hat{x}_2$.

Property 3. In a latent variable model $\mathcal{M}(e, g, p_X)$ that disentangles a causal process of type shown in Figure 1, the only causal feature of $\hat{x}$ w.r.t. generative factor G_i is $\mathbf{Z}_I \forall i$.

We now propose two evaluation metrics in the next section that are consequences of Properties 2 and 3. To study the disentanglement of a causal process of the type shown in Figure 1, we need datasets that reflect the generative process, and we hence introduce one in Section which offers several advantages such as realism, semantic confounding and complex backgrounds over existing datasets in addition to being generated from a two-level causal graph with confounders.

Evaluation Metrics

For causal disentanglement, the encoder e of a model $\mathcal{M}$ (e, g, p_X) should learn the *mapping* from G_i to $\mathbf{Z}_I$ without any influence from confounding in the data distribution p_X. (This would be equivalent to marginalizing over the confounder while computing direct causal effect between two variables.) If a model is able to map each G_i to a unique $\mathbf{Z}_I$, we say that the learned latent space $\mathbf{Z}$ is unconfounded. We call this property as *Unconfoundedness* (UC). UC captures the essentials of Property 2 as it relies on the mapping between G_i and $\mathbf{Z}_I$.

Secondly, when the latent space is unconfounded, a counterfactual instance of x w.r.t. generative factor G_i, x_I^{cf} (i.e., the counterfactual of x with change in only G_i) can be generated by intervening on the latents of x corresponding to G_i, $\mathbf{Z}_I^x$ and any change in the latent dimensions of $\mathbf{Z}$ that are not responsible for generating G_i, i.e. $\mathbf{Z}_{\setminus I}^x$, should have no influence on the generated counterfactual instance x_I^{cf} w.r.t. generative factor G_i. We call this property as *Counterfactual Generativeness* (CG). To explain with an example, consider an image of an ball in a certain background. The CG metric emphasises the fact that *"intervening on the latents corresponding to the background should only change the background and intervening on the latents corresponding to texture or shape of the ball should not change the background"*. Thus, CG follows from Property 3 as it is based on the fact that only causal effect on x_I^{cf} w.r.t. generative factor G_i is from $\mathbf{Z}_I^x$. We now formally define the two metrics.

Unconfoundedness (UC) Metric

The UC metric evaluates how well distinct generative factors G_i are captured by distinct sets of latent dimensions $\mathbf{Z}_I$ with no overlap (Figure 2). If a model encodes the underlying generative factor G_i of an instance x as a set of latent dimensions $\mathbf{Z}_I^x$, we define UC measure as:

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

where $S = \binom{n}{2}$ is the number of pairs of generative factors $(G_i, G_j), i \neq j$. We are in effect, finding the *Jaccard similarity coefficient* among all possible pairs of latent variables corresponding to different (G_i, G_j) to know how each pair of (G_i, G_j) are captured by unconfounded latent dimensions. To find correspondences between $\mathbf{Z}_I$ and G_i, we can use any existing metrics like (Suter et al. 2019; Chen et al. 2018) but we use the IRS measure (Suter et al. 2019) as it works on principles of interventions and is grounded on the properties of a disentangled causal process. For each generative factor G_i, IRS finds latents Z_I that are robust to interventions to G_j; $j \neq i$. If all generative factors are disentangled into distinct sets of latent factors, we get a UC score of 1. If all generative factors share the same set of latent factors, we get a UC score of 0. This definition of UC metric can be generalized to also check for unconfoundedness of multiple generative factors at a time.

Metrics closest to UC are MIG (Chen et al. 2018) and DCI (Eastwood and Williams 2018). Even though MIG penalizes non-axis aligned representations, it does not consider the case of multiple generative factors having the same latent representation, and hence may not capture unconfoundedness in a true sense. The Disentanglement(D) score in DCI uses correlation-based models to predict G_i given Z, and is hence not causal.

Counterfactual Generativeness (CG) Metric

When a latent variable model $\mathcal{M}$ achieves *unconfoundedness*, we can perform interventions on any specific $\mathbf{Z}_I$ to generate counterfactual instances without any confounding effect. That is, the generator g is able to generate counterfactual instances in a flexible and controlled manner. We call this *counterfactual generativeness*. In latent variable models that work on image datasets, to the best of our knowledge, this is the first effort to use generated images to *quantitatively* evaluate the level of disentanglement. To define CG metric mathematically, we need the notion of Average and Individual Causal Effect, which we provide below.

Definition 2. (Average Causal Effect). The Average Causal Effect (ACE) of a random variable Z on a random variable X for a treatment $do(Z) = \alpha$ with reference to a baseline treatment $do(Z) = \alpha^*$ is defined as $ACE^X_{do(Z=\alpha)} := \mathbb{E}[X|do(Z = \alpha)] - \mathbb{E}[X|do(Z = \alpha^*)]$.

Individual Causal Effect (ICE) can be defined similar to Definition 2 by replacing the expectation with probability as $ICE^x_{do(Z=\alpha)} := p[x|do(Z = \alpha)] - p[x|do(Z = \alpha^*)]$. Perfect disentanglement makes the generative model satisfy the positivity assumption (Hernan and Robins 2019) and allows us to approximate ACE with mean of ICEs taken over the dataset. Based on the above definitions, our *counterfactual generativeness* (CG) metric is defined as:

$$CG := \mathbb{E}_I[|ACE^{X_I^{cf}}_{\mathbf{Z}_I^x} - ACE^{X_{\setminus I}^{cf}}_{\mathbf{Z}_{\setminus I}^x}|] \tag{2}$$

$ACE^{X_I^{cf}}_{\mathbf{Z}_I^x}$ and $ACE^{X_{\setminus I}^{cf}}_{\mathbf{Z}_{\setminus I}^x}$ are defined to be the average causal effects of $\mathbf{Z}_I^x$ and $\mathbf{Z}_{\setminus I}^x$ on the respective counterfactual quantities x_I^{cf} and $x_{\setminus I}^{cf}$ (recall that $I \subset \{1, 2, \dots, m\}$ denotes the set of indices among the latent factors learned in the model that correspond to the G_i^{th} generative factor). So, the CG metric calculates the normalized sum of differences of average causal effects of $\mathbf{Z}_I^x$ and $\mathbf{Z}_{\setminus I}^x$ on the generated counterfactual quantities w.r.t. G_i (recall that for causal disentanglement, only $\mathbf{Z}_I^x$ should have causal effect on x_I^{cf} w.r.t. generative factor G_i; recall the ball, background example). Since counterfactual outcomes with respect to a model can be generated through interventions, we approximate ACE with the average of ICEs taken over the empirical distribution p_X. The

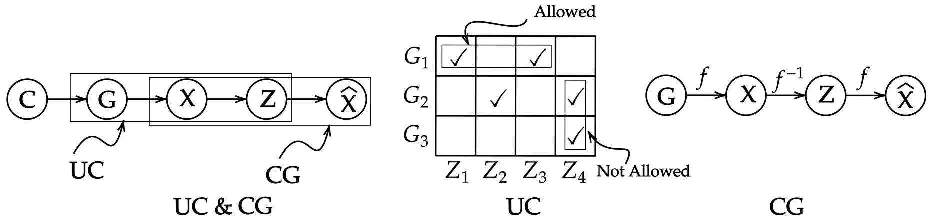

Figure 2: Left: UC metric relates G, X and Z. CG metric relates X, Z and $\hat{X}$. Center: According to UC metric, in a model $\mathcal{M}, G_1$ is allowed to be captured by Z_1, Z_3 but it is not allowed for Z_4 to capture both G_2, G_3 (this would suggest confounding). Right: Generative factors G generate image x through an unknown causal mechanism f, our goal in learning a disentangled representation is to learn f^{-1} and hence f that transforms observation x into latent dimensions Z and latent dimensions to reconstruction $\hat{X}$.

practical version of the CG metric is hence:

$$CG := \mathbb{E}_I \left[|ACE^{X_I^{cf}}_{\mathbf{Z}_I^X} - ACE^{X_{\setminus I}^{cf}}_{\mathbf{Z}_{\setminus I}^X}| \right] \approx \mathbb{E}_I \left[|\mathbb{E}_{x \sim p_x}[ICE^{x_I^{cf}}_{\mathbf{Z}_I^x} - ICE^{x_{\setminus I}^{cf}}_{\mathbf{Z}_{\setminus I}^x}]| \right]$$
$$\approx \frac{1}{n} \left[|\frac{1}{L}[ICE^{x_I^{cf}}_{\mathbf{Z}_I^x} - ICE^{x_{\setminus I}^{cf}}_{\mathbf{Z}_{\setminus I}^x}]| \right] \tag{3}$$

where L is the size of the dataset. Definition 2 holds for *real* random variables, but in latent variable models, x_I^{cf} is an image on which there is no clear way of defining causal effect of latents. Extending the notations, let G_{ik}^x represent the k^{th} value taken by i^{th} generative factor for a specific image x (e.g., if i = shape of an object, then k = cone). For this work, we define $ICE^{x_I^{cf}}_{\mathbf{Z}_I^x}$ to be the difference in prediction probability (of a pre-trained classifier) of G_{ik}^x given the counterfactual image x_I^{cf} generated when $do(\mathbf{Z}_I^x = \mathbf{Z}_I^x)$ (i.e. no change in latents of current instance) and when $do(\mathbf{Z}_I^x = baseline(\mathbf{Z}_I^x))$. Mathematically,

$$\begin{aligned} ICE^{x_I^{cf}}_{\mathbf{Z}_I^x} &:= |p(G_{ik}^x | x_I^{cf}, do(\mathbf{Z}_I^x = \mathbf{Z}_I^x)) \\ &\quad - p(G_{ik}^x | x_I^{cf}, do(\mathbf{Z}_I^x = baseline(\mathbf{Z}_I^x))| \end{aligned} \tag{4}$$

$$\begin{aligned} ICE^{x_{\setminus I}^{cf}}_{\mathbf{Z}_{\setminus I}^x} &:= |p(G_{ik}^x | x_{\setminus I}^{cf}, do(\mathbf{Z}_{\setminus I}^x = \mathbf{Z}_{\setminus I}^x)) \\ &\quad - p(G_{ik}^x | x_{\setminus I}^{cf}, do(\mathbf{Z}_{\setminus I}^x = baseline(\mathbf{Z}_{\setminus I}^x))| \end{aligned} \tag{5}$$

We use $baseline(\mathbf{Z}_I^x)$ as the latent dimensions that are maximally deviated from the current latent values $\mathbf{Z}_I^x$ (taken over the dataset) to ensure that we get a reasonably different image from the current image x w.r.t. generative factor G_i. $baseline(\mathbf{Z}_I^x)$ can be 0 or $\mathbb{E}_{x \sim p_X}(\mathbf{Z}_I^x)$ depending on the dataset and application. In the ideal scenario, Equation 4 is expected to output 1 because $\mathbf{Z}_I^x$ is the only causal feature of G_{ik}^x. Equation 5 is expected to output 0 because $\mathbf{Z}_{\setminus I}^x$ is not causally responsible for generating G_{ik}^x. Now it is easy to see that, for causal disentanglement, CG score in Equation 3 is 1; and for poor disentanglement, CG score is 0. The proposed UC and CG metrics can also be used irrespective of presence of confounders in the data generating process. The algorithms detailing the implementation of UC and CG metrics are provided in the Appendix.

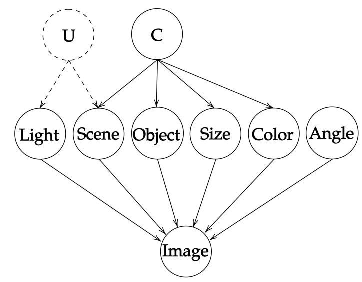

Figure 3: Image generating process of CANDLE

Dataset

To study causally disentangled representations, we introduce an image dataset called CANDLE (Causal ANalysis in DisentangLed rEpresentations) with 6 data generating factors along with both observed and unobserved confounders. Its generation follows the causal directed acyclic graph shown in Figure 3 which resembles our setting of causal graphs introduced in Figure 1. During generation, the *Image* has influences from confounders **U** (unobserved), and **C** (observed) through intermediate generative factors such as *Object* and *size*. It contains observed confounding in the form of semantic constraints such as overly large objects not being in indoor scenes (full list in Appendix). Unobserved confounding shows up in the interaction between the artificial light source and the scene's natural lighting conditions as it interacts with the foreground object producing shadows. Another source of subtle confounding in the dataset is how the location of the object and its size are confounded by *depth*, where a larger object that is farther-off and a smaller object nearby occupy the same pixel real-estate in the image, explored in (Träuble et al. 2021). Sample images from CANDLE are shown in figure 4 and a comparison with existing datasets used commonly in disentangled representation learning is provided in the Appendix where we also highlight the important features that are unique to CANDLE compared to existing datasets.

Dataset Creation. CANDLE is generated using Blender (Community 2018), a free and open-source 3D computer graphics suite which allows for manipulating background high-dynamic range images (HDRI images)

Figure 4: Sample images from CANDLE. Different objects appear in different colors, shapes, rotations and in different backgrounds respecting the causal graph in Figure 3

and adding foreground elements that inherit the natural light of the background. Foreground elements also naturally cast shadows and this greatly increases the realism of the dataset while allowing for it to be simulated for ease of creation. Since high-quality HDRI images are easy to obtain, it allows for multiple realistic backgrounds, unlike the plain colors in dSprites (Matthey et al. 2017) or colored strips in MPI3D (Gondal et al. 2019). Having complex backgrounds and the position of the foreground object varying between images adds another level of complexity while modeling the dataset for any downstream task. Having specific objects of interest in representation learning tasks puts more responsibility on the models being learned on the dataset to reconstruct images that do not leave out small objects in the reconstruction. To aid such reconstructions, bounding boxes of foreground objects are included in CANDLE's metadata. To further help with realism, we ensure that the capturing camera was not kept stationary and produced a fair amount of random jitter. At every stage, the dataset is made in such a way that extensions to it by adding objects or modifying the background scene is trivial (see Appendix for details).

CANDLE consists of $12,546$ images as 320×240 images and corresponding JSON files containing the factors of variation of each image (samples included in supplementary material). Recent works in disentangled representation learning have focused on identifying causal relationships in latent representations and the causal effects of latent representations on outcome variables (Yang et al. 2020; Chattopadhyay et al. 2019), which our dataset can readily support due to availability of the required ground truth. Another use case of CANDLE would be in counterfactual generation algorithms (Chang et al. 2018; Goyal et al. 2019a; Ilse et al. 2020), which we leave for future work.

Details of Factors of Variation. Background scenes of CANDLE are panoramic HDRI images of 4000×2000 resolution for accurate reproduction of the scene's lights on the object. Foreground objects are placed on the floor (without complete occlusion to guarantee presence of every label in the image). Objects are sized for semantic correctness in relation to the background (e.g., juxtaposing a very large cube and a building is unrealistic). Care is taken to make sure that significant overlap between objects and the background is eliminated. An artificial light source is added to the scene which also casts shadows in 3 positions - left, middle (overhead) and right. This is an unobserved confounding variable in the sense that it could conflict with the scene's illumination. The light source is kept invariant across all objects in the image i.e., the light's position is the same irrespective of other object variables. The rotations of foreground objects are in the vertical axis. This variable is specifically chosen as it has visible differences in a subset of objects but may be interpreted as noise in the rest. For more details on CANDLE, please see Appendix. We empirically observe that when the object of interest is small in the image and the image contains significant variations in the background scene, unlike on datasets such as MPI3D (Gondal et al. 2019) where foreground object is small but background is black/plain, reconstructions by standard latent variable models tend to not retain the foreground objects. One can use high multiplicative factors for the reconstruction term in the learned objective function, but this leads to bad latent representations (Kim and Mnih 2018). We show how the bounding box information provided in CANDLE's metadata is used as weak supervision to solve this problem partially in Section .

Learning Disentangled Representations Using Weak Supervision

We now provide a simple methodology to improve over existing models that learn disentangled representations by using the bounding box-level supervision information in CANDLE. Since there is a known trade-off between reconstruction quality and disentanglement in VAE-based models(Kim and Mnih 2018), instead of giving high weightage to reconstruction quality during training at the cost of worse disentanglement, we hypothesize that paying more attention to the quality of reconstructions of specific foreground objects whose bounding box is known provides a more favorable trade-off between reconstructions and disentanglement. We improve the existing semi-supervised Factor-VAE(Kim and Mnih 2018) loss with an additional loss term that weights regions in the bounding box higher than others to aid in better reconstructions of foreground objects. We call this method *Semi-Supervised Factor-VAE with additional Bounding Box supervision* or *SS-FVAE-BB*. Our loss function w.r.t. dataset $\mathcal{D} = \{x_i\}_{i=1}^{L}$ is given by:

$$\mathcal{L}_{SS-FVAE-BB} = \mathcal{L}_{(Factor-VAE)} + \lambda \sum_{i=1}^{L} ||x_i \odot w_i - \hat{x}_i \odot w_i||_2^2 \quad (6)$$

where $w_i \in \{0, 1\}^{320 \times 240 \times 3}$ is an indicator tensor with 1s in the region of the bounding box and 0s elsewhere, λ is a hyperparameter and $\odot$ is the Hadamard (elementwise) product. Our experimental results (Table 1) show that the proposed method improves *UC* score while matching the best *CG* score achieved by state-of-the-art models. *SS-FVAE-BB* can also be used with the datasets without bounding box information by using any segmentation techniques that highlight the objects of interest in the images.

Experimental Results

To study causal disentanglement, we performed experiments on well-known unsupervised disentanglement methods as well as their corresponding semi-supervised variants: β-VAE (Higgins et al. 2017), β-TCVAE (Chen et al. 2018), DIP-VAE (Kumar, Sattigeri, and Balakrishnan 2017), and Factor-VAE (Kim and Mnih 2018) using the proposed dataset and evaluation metrics. We also included studies on other existing datasets – dSprites, MPI3D-Toy, and a synthetic toy dataset with extreme confounding – for completeness of analysis and comparison. The learned models are compared using IRS, $DCI(D)$, UC and CG metrics. We use the open-source disentanglement library (Locatello et al. 2019) for training models. Semi-supervision is provided by using labels for 10% of data points. Additional details on the experimental setup and qualitative results are provided in the Appendix. In the results below, ρ refers to the number of latent dimensions that we choose to attribute for each generative factor.

Model	*IRS*	*DCI* (D)	*UC* $\rho = 5$	*CG* $\rho = 5$	*UC* $\rho = 7$	*CG* $\rho = 7$
β-VAE	0.85	0.18	0.11	0.24	0.08	0.22
β-TCVAE	0.82	0.10	0.11	0.25	0.08	0.25
DIP-VAE	0.33	0.08	0.11	0.21	0.15	0.22
Factor-VAE	**0.88**	0.15	0.13	0.26	0.08	**0.28**
SS-β-VAE	0.74	**0.18**	0.11	**0.28**	0.08	0.19
SS-β-TCVAE	0.68	0.17	0.11	0.23	0.08	0.19
SS-DIP-VAE	0.35	0.08	0.11	0.22	0.15	0.22
SS-Factor-VAE	0.61	0.16	0.24	**0.28**	0.14	0.22
SS-FVAE-BB	0.61	0.13	**0.27**	**0.28**	**0.18**	**0.28**

Table 1: Results on CANDLE

Model	*IRS*	*DCI* (D)	*UC* $\rho = 1$	*CG* $\rho = 1$	*UC* $\rho = 2$	*CG* $\rho = 2$
β-VAE	0.49	0.16	0.70	0.12	0.46	0.10
β-TCVAE	**0.78**	0.43	**0.90**	**0.19**	0.60	**0.19**
DIP-VAE	0.12	0.03	**0.90**	0.04	0.60	0.03
Factor-VAE	0.44	0.13	**0.90**	0.07	0.60	0.06
SS-β-VAE	0.52	0.23	**0.90**	0.17	0.60	0.17
SS-β-TCVAE	0.72	**0.50**	**0.90**	0.18	**0.67**	0.18
SS-DIP-VAE	0.20	0.04	0.40	0.08	0.13	0.06
SS-Factor-VAE	0.47	0.19	**0.90**	0.15	0.33	0.14

Table 2: Results on dSprites

Results on CANDLE: Table 1 shows the results of different performance metrics, including the proposed UC and CG metrics, when the considered generative models are learned on the CANDLE dataset (the '*SS-*' prefix refers to the 'Semi-Supervised' variants). The table shows low UC and CG scores in general, motivating the need for better disentanglement methods. Owing to the complex background, models find it difficult to reconstruct foreground objects during the learning process which causes the learned latent dimensions corresponding to foreground objects difficult to identify. Changing the value of ρ has its consequences. We observe higher (but not high enough for good disentanglement) UC scores when $\rho = 5$. However, when $\rho = 7$, we observe low UC scores because multiple latent dimensions are confounded. Owing to the complex background, models learn to reconstruct images with little to no information about the foreground object which also leads to low CG scores. Much of the observed CG score can be attributed to the *scene* factor because scenes are reconstructed well (see Appendix). Introducing weak supervision in the training of the generative model using our proposed method *SS-FVAE-BB* with $\lambda = 2$ improves UC score without compromising CG score.

Model	*IRS*	*DCI* (D)	*UC* $\rho = 1$	*CG* $\rho = 1$	*UC* $\rho = 2$	*CG* $\rho = 2$
β-VAE	0.57	0.23	0.52	0.10	0.34	0.12
β-TCVAE	0.57	0.22	0.52	0.12	0.35	0.14
DIP-VAE	0.22	0.23	0.28	0.10	0.19	0.14
Factor-VAE	0.52	**0.34**	0.71	**0.14**	0.47	**0.16**
SS-β-VAE	0.60	0.28	**0.80**	0.10	**0.67**	0.09
SS-β-TCVAE	**0.64**	0.26	**0.80**	0.09	**0.67**	0.15
SS-DIP-VAE	0.35	0.25	0.52	0.10	0.34	0.11
SS-Factor-VAE	0.56	0.30	**0.80**	0.12	**0.67**	0.14

Table 3: Results on MPI3D-Toy dataset

Results on dSprites & MPI3D-Toy: For completeness of analysis, we conducted experiments on training the above-mentioned generative models on existing datasets with no confounding like dSprites & MPI3D (Tables 2, 3). The UC and CG metrics can be used to evaluate models under this setting too. As we are training models on full datasets without any observable confounding effect, we observe high UC scores when $\rho = 1$. However, when $\rho = 2$, results start to show limitations of existing models to disentangle completely. Additional results on confounded versions of dSprites and MPI3D-Toy datasets, as well as on a synthetic toy dataset with confounding that we created for purposes of analysis, are deferred to the Appendix owing to space constraints. The results in general show that there is no single model that outperforms w.r.t. all the metrics, which shows the importance of datasets like CANDLE and evaluation metrics, such as UC and CG scores developed using the principles of causality, to uncover sources of bias that were not considered previously.

Conclusions

A causal view of disentangled representations is important for learning trustworthy and transferable mechanisms from one domain to another . We build on the very little work along this direction by analysing the properties of causal disentanglement in latent variable models. We propose two evaluation metrics and a dataset which are used to uncover the causal disentanglement in existing disentanglement methods. We also improved over existing models by introducing a simple weakly supervised disentanglement method. We hope that newer machine learning models benefit from our metrics and dataset in developing causally disentangled representation learners.

Acknowledgments

We are grateful to the Ministry of Education, India for the financial support of this work through the Prime Minister's Research Fellowship (PMRF) and UAY programs. This work has also been partly supported by Honeywell and a Google Research Scholar Award, whom we are thankful to. We thank the anonymous reviewers for their valuable feedback that helped improve the presentation of this work.

References

Bengio, Y.; Courville, A.; and Vincent, P. 2013. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8): 1798–1828.

Burgess, C.; and Kim, H. 2018. 3D Shapes Dataset. https://github.com/deepmind/3dshapes-dataset/. Visited on 2020-10-25.

Chang, C.-H.; Creager, E.; Goldenberg, A.; and Duvenaud, D. 2018. Explaining image classifiers by counterfactual generation. *arXiv preprint arXiv:1807.08024*.

Chattopadhyay, A.; Manupriya, P.; Sarkar, A.; and Balasubramanian, V. N. 2019. Neural Network Attributions: A Causal Perspective. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, 981–990. Long Beach, California, USA: PMLR.

Chen, J.; and Batmanghelich, K. 2020. Weakly Supervised Disentanglement by Pairwise Similarities. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04): 3495–3502.

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

Community, B. O. 2018. Blender - a 3D modelling and rendering package. http://www.blender.org. Visited on 01-08-2020.

Creager, E.; Madras, D.; Jacobsen, J.-H.; Weis, M.; Swersky, K.; Pitassi, T.; and Zemel, R. 2019. Flexibly fair representation learning by disentanglement. In *International conference on machine learning*, 1436–1445. PMLR.

Dittadi, A.; Träuble, F.; Locatello, F.; Wuthrich, M.; Agrawal, V.; Winther, O.; Bauer, S.; and Schölkopf, B. 2021. On the Transfer of Disentangled Representations in Realistic Settings. In *International Conference on Learning Representations*.

Eastwood, C.; and Williams, C. K. 2018. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*.

Fidler, S.; Dickinson, S.; and Urtasun, R. 2012. 3D Object Detection and Viewpoint Estimation with a Deformable 3D Cuboid Model. In Pereira, F.; Burges, C. J. C.; Bottou, L.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Gilpin, L. H.; Bau, D.; Yuan, B. Z.; Bajwa, A.; Specter, M.; and Kagal, L. 2018. Explaining Explanations: An Approach to Evaluating Interpretability of Machine Learning. *CoRR*, abs/1806.00069.

Gondal, M. W.; Wuthrich, M.; Miladinovic, D.; Locatello, F.; Breidt, M.; Volchkov, V.; Akpo, J.; Bachem, O.; Schölkopf, B.; and Bauer, S. 2019. On the Transfer of Inductive Bias from Simulation to the Real World: a New Disentanglement Dataset. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Goyal, Y.; Feder, A.; Shalit, U.; and Kim, B. 2019a. Explaining classifiers with causal concept effect (cace). *arXiv preprint arXiv:1907.07165*.

Goyal, Y.; Wu, Z.; Ernst, J.; Batra, D.; Parikh, D.; and Lee, S. 2019b. Counterfactual visual explanations. *arXiv preprint arXiv:1904.07451*.

Hernan, M.; and Robins, J. 2019. *Causal Inference*. Chapman & Hall/CRC Monographs on Statistics & Applied Probab. Taylor & Francis. ISBN 9781420076165.

Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M. M.; Mohamed, S.; and Lerchner, A. 2017. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *ICLR*.

Ilse, M.; Tomczak, J. M.; Louizos, C.; and Welling, M. 2020. Diva: Domain invariant variational autoencoders. In *Medical Imaging with Deep Learning*, 322–348. PMLR.

Janzing, D. 2019. Causal regularization. In *Advances in Neural Information Processing Systems*, 12704–12714.

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

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kocaoglu, M.; Snyder, C.; Dimakis, A. G.; and Vishwanath, S. 2018. CausalGAN: Learning Causal Implicit Generative Models with Adversarial Training. In *International Conference on Learning Representations*.

Kumar, A.; Sattigeri, P.; and Balakrishnan, A. 2017. Variational inference of disentangled latent concepts from unlabeled observations. *arXiv preprint arXiv:1711.00848*.

LeCun, Y.; Huang, F. J.; and Bottou, L. 2004. Learning methods for generic object recognition with invariance to pose and lighting. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, II–104. IEEE.

Lipton, Z. C. 2018. The mythos of model interpretability. *Queue*, 16(3): 31–57.

Locatello, F.; Bauer, S.; Lucic, M.; Raetsch, G.; Gelly, S.; Schölkopf, B.; and Bachem, O. 2019. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, 4114–4124.

Locatello, F.; Poole, B.; Rätsch, G.; Schölkopf, B.; Bachem, O.; and Tschannen, M. 2020. Weakly-Supervised Disentanglement Without Compromises. *arXiv preprint arXiv:2002.02886*.

Matthey, L.; Higgins, I.; Hassabis, D.; and Lerchner, A. 2017. dSprites: Disentanglement testing Sprites dataset. https://github.com/deepmind/dsprites-dataset/. Visited on 2020-08-01.

Montero, M. L.; Ludwig, C. J.; Costa, R. P.; Malhotra, G.; and Bowers, J. 2021. The role of Disentanglement in Generalisation. In *International Conference on Learning Representations*.

Nie, W.; Karras, T.; Garg, A.; Debnath, S.; Patney, A.; Patel, A.; and Anandkumar, A. 2020. Semi-Supervised StyleGAN for Disentanglement Learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 7360–7369. PMLR.

O' Shaughnessy, M.; Canal, G.; Connor, M.; Rozell, C.; and Davenport, M. 2020. Generative causal explanations of black-box classifiers. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M. F.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 5453–5467. Curran Associates, Inc.

Pearl, J. 2009. *Causality*. Cambridge University Press. ISBN 9781139643986.

Peters, J.; Janzing, D.; and Schölkopf, B. 2017. *Elements of Causal Inference: Foundations and Learning Algorithms*. Adaptive Computation and Machine Learning series. Mass. ISBN 9780262037310.

Pitis, S.; Creager, E.; and Garg, A. 2020. Counterfactual Data Augmentation using Locally Factored Dynamics. *arXiv preprint arXiv:2007.02863*.

Rezende, D. J.; Mohamed, S.; and Wierstra, D. 2014. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. In Xing, E. P.; and Jebara, T., eds., *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, 1278–1286. Bejing, China: PMLR.

Ridgeway, K.; and Mozer, M. C. 2018. Learning deep disentangled embeddings with the f-statistic loss. In *Advances in Neural Information Processing Systems*, 185–194.

Schölkopf, B.; Locatello, F.; Bauer, S.; Ke, N. R.; Kalchbrenner, N.; Goyal, A.; and Bengio, Y. 2021. Toward Causal Representation Learning. *Proceedings of the IEEE*, 109: 612–634.

Shu, R.; Chen, Y.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Weakly Supervised Disentanglement with Guarantees. In *International Conference on Learning Representations*.

Suter, R.; Miladinovic, D.; Schölkopf, B.; and Bauer, S. 2019. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. In *International Conference on Machine Learning*, 6056–6065. PMLR.

Träuble, F.; Creager, E.; Kilbertus, N.; Locatello, F.; Dittadi, A.; Goyal, A.; Schölkopf, B.; and Bauer, S. 2021. On Disentangled Representations Learned from Correlated Data. In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 10401–10412. PMLR.

Yang, M.; Liu, F.; Chen, Z.; Shen, X.; Hao, J.; and Wang, J. 2020. CausalVAE: Structured Causal Disentanglement in Variational Autoencoder. *arXiv preprint arXiv:2004.08697*.

Zhu, S.; Ng, I.; and Chen, Z. 2020. Causal Discovery with Reinforcement Learning. In *International Conference on Learning Representations*.

Zhu, Y.; Min, M. R.; Kadav, A.; and Graf, H. P. 2020. S3VAE: Self-Supervised Sequential VAE for Representation Disentanglement and Data Generation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6537–6546.

Zmigrod, R.; Mielke, S. J.; Wallach, H.; and Cotterell, R. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. *arXiv preprint arXiv:1906.04571*.