CANDLE: An Image Dataset for Causal Analysis in Disentangled Representations

Abbavaram Gowtham Reddy
IIT Hyderabad, India
cs19resch11002@iith.ac.in

Benin Godfrey L
IIT Hyderabad, India
benin.godfrey@cse.iith.ac.in

Vineeth N Balasubramanian
IIT Hyderabad, India
vineethnb@iith.ac.in

Abstract

Confounding effects are inevitable in real-world observations. It is useful to know how the data looks like without confounding. Coming up with methods that identify and remove confounding effects are important for various downstream tasks like classification, counterfactual data augmentation, etc. We develop an image dataset for ***C****ausal* ***AN****alysis in* ***D****isentang****L****ed r****E****presentations(CANDLE). We also propose two metrics to measure the level of disentanglement achieved by any model under confounding effects. We empirically analyze the disentanglement capabilities of existing methods on dSprites and CANDLE datasets.*

1. Introduction

Ground-truth causal information is needed for evaluating models that learn causal effects from observational data [11]. Real-world datasets for such tasks are hard to obtain because of *the fundamental problem of causal inference* and the infeasibility of randomized control trials. Therefore, we are constrained to use synthetic datasets and few controlled real-world datasets [5], most of which are non-image datasets. Also, synthetic datasets are hardly realistic and often favor granularity of variations rather than semantic constraints imposed by nature. We seek to fill the gap between semantic realism of typical image datasets and ground-truth data generation of simulated datasets. We develop a realistic, but simulated image dataset(Figure 1) for **C**ausal **AN**alysis in **D**isentang**L**ed r**E**presentations(CANDLE), whose generation follows a causal graph with unobserved confounding effects, mimicking the real world.

We ask counterfactual questions on CANDLE to evaluate a model's ability to learn the underlying causal structure. In few applications, generative factors are correlated due to physical properties of nature(e.g., shadow and light position). But in many applications it is helpful to isolate the generative factors completely to understand various properties of learned representations. Models that learn disentangled representations from confounded datasets are useful for learning representations from limited examples. Many

Figure 1: Sample images from CANDLE.

current models either assume confounding is not present, or ignore it even if it is. We encourage models that consider confounding by creating a dataset with both observed and unobserved confounding effects with complex background. The motivation to create a dataset in this fashion arose from recent ideas requiring weak supervision for reliable representation learning [9]. We also propose two metrics to evaluate the level of disentanglement under confounding. Existing metrics use generative factors and latent representations to come up with disentanglement score. Our metrics concentrate on the entire process of data generation, latent representations and reconstructions. Our metrics are based on the principles of causality and develops on the very little work along this direction [14].

2. Related Work

There are several datasets available for causal inference and causal discovery [5] but they are non image datasets. Image datasets generated for causal analysis are studied in disentangled representation learning(e.g., dSprites [12], MPI3D [4]). But CANDLE is unique such that it has complex backgrounds and object's position is not fixed across images making it challenging for models to learn disentangled representations. Currently various disentanglement metrics are available(e.g., MIG [1], DCI [3], IRS [14]), but they rely on the learned latent space to evaluate the disentangled representations and the effects of confounding are not considered. IRS [14] considers the confounding effect to be present in the data generation mechanism but its implications were not considered while formulating the metric(e.g., two generative factors that are correlated can be en-

coded by a single latent factor but we like to penalise confounded encodings). Since we are interested in counterfactual questions on confounded datasets, reconstructions are important.

3. The CANDLE Dataset

Causal DAG $\mathcal{G}$, based on which the dataset is generated is shown in Figure 2. The dataset is generated usings Blender [2], which allows for manipulating the background HDRI images and adding foreground elements that inherit the natural lighting of the background.

Foreground elements naturally cast shadows to interact with the background. This greatly increases the realism of the dataset while allowing for it to remain simulated. Unlike other datasets, the position of the foreground object varies between images. This adds another level of hardness for any downstream task. Having specific objects of interest in disentanglement representation learning puts more responsibility to reconstruct images that do not leave out small objects in reconstruction and we use this aspect in the design of metrics. We also ensured that the capturing camera was not kept stationary and produced a fair amount of random jitter. Each variation of possible values is rendered out as one instance in the dataset to form a fully specified system. The Unobserved confounder is implicit in between light source and scene, observed confounding can be obtained by selecting the part of the dataset as desired. The nodes and corresponding values taken by the nodes in our causal graph are shown in Table 1. CANDLE aims to augment and alleviate the transferability concerns by using high-resolution panoramic HDRI images of real-world scenes which can be manipulated by standard 3D modelling to produce a realistic but simulated dataset and corresponding ground truth data. This ground-truth data, though impossible to obtain in real-life, provide valuable information which can be exploited in various metrics of disentanglement. The dataset consists of 4050 images as 320×240 png images and corresponding JSON files containing the factors of variation of each image. Background scenes are panoramic HDRI images of 4000x2000 resolution for accurate reproduction of the scene's lights on the object. To simplify the dataset, the objects are placed on the floor with no complete occlusion to guarantee presence of every label in the image. Objects are small for semantic correctness in relation to the background. Care is taken to make sure that significant overlapping between the 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 a confounding variable in the sense that it conflicts with the scene's shadow and also is invariant across all objects in the image. Rotations of objects are in the vertical axis.

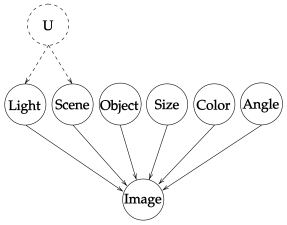

Figure 2: Data generating DAG with unobserved confounder(U).

Concepts	Values
Object	Cube, Sphere, Cylinder, Cone, Torus
Color	Red, Blue, Yellow, Purple, Orange
Size	Small, Medium, Large
Rotation	$0°, 15°, 30°, 45°, 60°, 90°$
Scene	Indoor, Playground, Outdoor
Light	Left, Middle, Right

Table 1: factors of variation

Weak supervision to disentangling models can be provided in a variety of ways. This variety of supervisions are provided natively with CANDLE . This allows any current and future model to use supervision for learning and evaluation as required. E.x. paired images as supervision to learn disentangled representations has been explored recently [10]. All such supervisions can be obtained by simple querying on the metadata of our dataset. We empirically observed that when the object of interest is small in the image and image contains much variation in other parts (e.g. background) unlike MPI3D [4] where foreground object is small but background is black/plain, reconstructions by standard VAE-based models tend to not retain the foreground objects. One option is to use high multiplicative factor for the reconstruction term but it may lead to bad latent representations[7]. Using the bounding boxes as supervision, we can give more importance to the area given by bounding boxes in reconstruction loss to create a better trade-off between reconstruction and latent representations.

4. Evaluation Metrics

We propose two metrics to evaluate disentangled representation learners. Let G_i be the i^{th} generative factor and G_{ik} be the k^{th} value taken by i^{th} generative factor (e.g., $i \sim$ shape, $k \sim$ cone). Let $Z_j \in Z$ be the j^{th} latent dimension and M, N be the number of latent dimensions and number of generative factors respectively. Let $I \subset \{1, 2, \ldots, M\}$ be subset of indices used to index latent dimensions. Let $\mathbb{D} = \{x_i\}_{i=1}^{L}$ denote dataset of images obtained by following $\mathcal{G}$. Let Z_I^x be the learned, indexed latent dimensions indexed by I for a specific image x and $Z_{\setminus I}^x$ be the learned, indexed latent dimensions indexed by the set $\{1, 2, \ldots, M\} - I$ for a specific image x.

Using the definition of *disentangled causal process* [14], irrespective of the presence of confounder U during data generating process, for perfect disentanglement, a model should learn the *independent causal mechanisms* relating generative factors and outcome variable without any influence from confounders U while intervening on any of the generative factors. If a model is able to disentangle a causal

process for an outcome variable, we say that learnt latent space is unconfounded. We call this phenomenon as *Unconfoundedness*(UC). When the latent space is unconfounded, counterfactuals can be generated without any confounding bias. We call this phenomenon as *Counterfactual Generativeness*(CG). So, we look at causal disentanglement as the combination of UC and CG.

4.1. Unconfoundedness(UC) Metric

The UC measure evaluates how well distinct generative factors G_i are captured by different sets of latent dimensions Z_I with no overlap. A seemingly related metric to UC is the DCI [3] metric. The D(disentanglement) score in DCI metric is similar to UC metric but our analysis is based on interventions and causal influences rather than correlation based models used for predicting G_i given Z. Also the way we compute UC metric is different (Equation 1). We assume that the variation in each G_i is captured by a set of latent dimensions Z_I because of the fact that more than one latent dimension can encode a single generative factor. For the simplicity of mathematical notation, we call Z_I^x as Q_i^x. Now, if a model captures the underlying generative factors G_i into a set of latent dimensions Q_i, we define UC measure as

$$UC \coloneqq 1 - \mathbb{E}_{x\sim\mathbb{D}}\big[\frac{1}{K}\sum_{i\neq j}\frac{|Q_i^x \cap Q_j^x|}{|Q_i^x \cup Q_j^x|}\big] \tag{1}$$

Where $K = \binom{N}{2}$, the number of all possible pairs of generative factors. We are essentially finding *Jaccard similarity coefficient* among all possible pairs of latent sets corresponding to different generative factors to know how each pair of generative factors are captured by unconfounded latent dimensions. If all the generative factors are disentangled into different sets of latent factors, we get $UC = 1$. In the worst case, if all generative factors share same set of latents, we get $UC = 0$. To find Z_I corresponding to G_i we use IRS measure [14] because it is closely related to our metric and it works on the principles of interventions.

4.2. Counterfactual Generativeness(CG) Metric

When G_i's are disentangled well, we can generate counterfactual images by intervening on any latent dimension in the generative model without worrying about confounding bias. To define CG metric, we use Average Causal Effect(ACE) and Individual Causal Effect(ICE)[13] of latents on generated images. Since latents in the generative model are at the root level, conditioning is same as intervening. This lets us to define *counterfactual generativeness* (CG) measure as

$$CG \coloneqq \mathbb{E}_I\big[|ACE_{Z_I}^{X_I^{cf}} - ACE_{Z_{\setminus I}}^{X_{\setminus I}^{cf}}|\big] \tag{2}$$

where X_I^{cf} represents the counterfactual image generated when latent factors Z_I are set to some interventional value. $ACE_{Z_I}^{X_I^{cf}}, ACE_{Z_{\setminus I}}^{X_{\setminus I}^{cf}}$ are defined to be the average causal effects of $Z_I, Z_{\setminus I}$ on the respective counterfactual images when the generative factor of interest is G_i. So, the CG measure calculates the normalized sum of differences of average causal effects of Z_I and $Z_{\setminus I}$ on the generated counterfactual images. Since counterfactual outcomes with respect to a model can be generated through interventions, we approximate ACE with average of ICEs over the dataset. Now CG metric is modified as follows.

$$CG \coloneqq \mathbb{E}_I\big[|ACE_{Z_I}^{X_I^{cf}} - ACE_{Z_{\setminus I}}^{X_{\setminus I}^{cf}}|\big] \tag{3}$$

$$\approx \frac{1}{N}\frac{1}{L}\big[|ICE_{Z_I^x}^{x_I^{cf}} - ICE_{Z_{\setminus I}^x}^{x_{\setminus I}^{cf}}|\big] \tag{4}$$

ACE definition [13] is for real random variables, but our target variable X_I^{cf} is an image, on which there is no clear way of defining causal effect of latents. For this work, we define $ICE_{Z_I^x}^{x_I^{cf}}$ to be the difference in prediction probability of G_{ik}(of a pre-trained classifier) given x_I^{cf} generated when $do(Z_I = Z_I^x)$(no change in latents of current example) and when $do(Z_I = baseline(Z_I^x))$. Mathematically,

$$\begin{aligned} ICE_{Z_I^x}^{x_I^{cf}} &= |P(G_{ik}|x_I^{cf}, do(Z_I = Z_I^x)) \\ &\quad - P(G_{ik}|x_I^{cf}, do(Z_I = baseline(Z_I))|\end{aligned} \tag{5}$$

We choose to use $baseline(Z_I^x) = max\_dev(Z_I^x, Z_I)$. Where $max\_dev(Z_I^x, Z_I)$ is the latent values that maximally deviated from from current latent values Z_I^x of an instance x(taken over dataset) to ensure that we get reasonably different image than current image w.r.t generative factor G_i. In the ideal scenario, $ICE_{Z_I^x}^{x_I^{cf}}$ is expected to output 1 because we are intervening on Z_I corresponding to G_i(which is the prediction outcome in the classification model) to generate counterfactual image. And $ICE_{Z_{\setminus I}^x}^{x_{\setminus I}^{cf}}$ is expected to output 0 because we are intervening on $Z_{\setminus I}$ that are not corresponding to G_i to generate counterfactual image. For perfect disentanglement, $CG = 1$. For poor disentanglement, $CG = 0$.

5. Experiments

We use the following state of the art models in disentanglement learning and corresponding semi supervised variants to evaluate their ability to disentangle using our metric on CANDLE and dSprites: β-VAE [6], DIP-VAE [8], Factor-VAE [7] and β-TCVAE [1]. Models are compared on IRS, UC and CG metrics.s ρ is the number of latent dimensions that we choose to attribute for each generative factor.

CANDLE & dSprites: We chose a subset of CANDLE to mimic observable confounding effect. Confounding used here is that few objects appear in only few colors: cube in red and blue, sphere in blue and yellow, cylinder in yellow and purple, cone in purple and orange and torus in orange and red. From the results it is evident that the models are getting low UC and CG scores which reveal the need for better disentangled methods under confounding scenarios. Observe the relatively high(but not high enough for good disentanglement) UC and CG scores when $\rho = 1$ but when $\rho = 2$, we observe low UC and CG scores because multiple latent dimensions are confounded and thus also affect reconstructions. Owing to the complex background, we observe that models learn to reconstruct images with little to no information about the foreground object which leads to low CG scores. In the table, 'SS' refers to 'Semi Supervised'. Accuracy of the trained model used in CG metric is 98.9%. We can also use UC and CG to evaluate the models that train on complete(no observed confounding) datasets like dSprites. As we are training models on the full dSprites dataset without any observable confounding effect, we observe high UC score only when $\rho = 1$(Table 2). The reason for low score in CG is that models are not able to generate reconstructions that capture difficult generative factors exactly like angle, position, etc. Accuracy of the trained model used in CG metric is 99.5%

Model	IRS	UC $\rho=1$	CG $\rho=1$	UC $\rho=2$	CG $\rho=2$
CANDLE					
β-VAE	**0.80**	0.40	0.21	0.13	0.20
β-TCVAE	0.50	0.70	0.20	0.29	0.20
DIP-VAE	0.24	0.40	0.13	0.29	0.14
Factor-VAE	0.47	0.40	0.22	0.13	0.15
SS-β-VAE	0.51	**0.80**	0.28	**0.60**	0.19
SS-β-TCVAE	0.49	**0.80**	0.25	0.46	**0.25**
SS-DIP-VAE	0.29	0.40	0.32	0.26	0.19
SS-Factor-VAE	0.46	0.40	**0.36**	**0.60**	0.21
dSprites					
β-VAE	0.40	**0.90**	0.12	**0.60**	0.10
β-TCVAE	0.62	**0.90**	0.12	**0.60**	0.13
DIP-VAE	0.48	0.80	0.10	0.53	0.10
Factor-VAE	0.52	**0.90**	0.14	**0.60**	0.14
SS-β-VAE	0.55	**0.90**	**0.16**	0.30	**0.15**
SS-β-TCVAE	**0.70**	**0.90**	0.15	0.33	0.14
SS-DIP-VAE	0.24	0.40	0.08	0.13	0.06
SS-Factor-VAE	0.47	**0.90**	0.15	0.33	0.14

Table 2: Comparison of IRS, UC and CG metrics on CANDLE & dSprites dataset for various models.

6. Conclusion

We introduced a dataset and two metrics to study the level of unconfoundedness achieved by a model. Despite repeated pivoting in the field of disentangled representation learning and the possible unification with causal factor learning or structure learning, we hope that this dataset and metrics help the community in providing more challenging and competing models in various tasks such as fairness, reasoning, evaluation etc.

References

[1] Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *NeurIPS*, 2018. 1, 3

[2] Blender Online Community. Blender - a 3d modelling and rendering package. http://www.blender.org. 2

[3] Cian Eastwood and Christopher KI Williams. A framework for the quantitative evaluation of disentangled representations. In *ICLR*, 2018. 1, 3

[4] Muhammad Waleed Gondal, Manuel Wuthrich, Djordje Miladinovic, Francesco Locatello, Martin Breidt, Valentin Volchkov, Joel Akpo, Olivier Bachem, Bernhard Schölkopf, and Stefan Bauer. On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset. In *NeurIPS*, 2019. 1, 2

[5] Ruocheng Guo, Lu Cheng, Jundong Li, P. Richard Hahn, and Huan Liu. A survey of learning causality with data:problems and methods. *arXiv preprint arXiv:1809.09337*, 2018. 1

[6] Irina Higgins, Loïc Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew M Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017. 3

[7] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. *arXiv preprint arXiv:1802.05983*, 2018. 2, 3

[8] Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. *arXiv preprint arXiv:1711.00848*, 2017. 3

[9] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *ICML*, 2019. 1

[10] Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. *arXiv preprint arXiv:2002.02886*, 2020. 2

[11] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *NeurIPS*, 2017. 1

[12] Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017. visited on 2020-08-01. 1

[13] J. Pearl. *Causality*. Cambridge University Press, 2009. 3

[14] Raphael Suter, Djordje Miladinovic, Bernhard Schölkopf, and Stefan Bauer. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. In *ICML*. PMLR, 2019. 1, 2, 3