Analyzing Representations through Interventions

Petra Poklukar[1,⋆][0000−0001−6920−5109], Michael C. Welle[1,⋆][0000−0003−3827−3824], Anastasiia Varava[1][0000−0002−0900−1523], and Danica Kragic[1][0000−0003−2965−2953]

Robotics, Perception and Learning Lab, EECS at KTH Royal Institute of Technology
{poklukar,mwelle,varava,dani}@kth.se

Abstract. Learning efficient and compact representations is crucial for various problems in artificial intelligence. Designing evaluation procedures that are independent from the task at hand is one of the key challenges of this area of research. Inspired by the theory of causal reasoning, we consider interventions for analysing data representation. Specifically, we leverage them to introduce *Interventional Score* for measuring disentanglement of the latent representations.

Keywords: Representation Learning · Interventions · Disentanglement Score.

1 Introduction and Related Work

The recent success of Deep Neural Networks in various domains has increased the need for more explainable and interpretable methods. While many recent works have been pursuing this direction [13], [16], [26], [21], difficulties already arise at the imprecise definition of "explainability" [14]. The advantage of many deep learning approaches is that the network can identify and make use of features that are automatically extracted from the data. In deep generative models such as Variational Autoencoders (VAE) [11], [19], features necessary to succeed at a given task are encoded in the low-dimensional latent space. Evaluating and understanding latent representations is important for their explainability and generality. One approach proposed in [2], and used in the subsequent works ([1], [3], [4], [25], [6]) is the notion of *disentanglement*. A representation is disentangled when one independent factor of variation or a underlying generative factor present in the data is associated with exactly one latent code [2]. Many recent methods encourage disentanglement by augmenting the loss function in VAEs [24].

While many explicit *disentanglement scores* have emerged ([9], [10], [12], [4], [5], [20], [22]) they are often heavily intertwined with the proposed method and it is therefore difficult to identify a universally *correct* score. Designing evaluation procedures for different models with respect to the quality of their representations rather than their performance on a specific tasks therefore remains an open question. The theory of causal reasoning, presented in [17], has recently gained more interest in the machine learning community [23], [22]. One concept that is of special interest in this work is the notion of an *intervention*. Interventions can be used to analyze causal dependencies between random variables as well as validate hypotheses about causal structures. This idea is explored in [22], where the authors use interventional sets to obtain a disentanglement score. The aforementioned scores are based on the assumption that a disentangled representation is the most desired and useful one for any given task. However, in [2] the

⋆ These authors contributed equally.

authors identified disentanglement as one of multiple *meta-priors*: disentanglement, hierarchical organization, semi-supervised learning, clustering structure. [24] provides an overview on how these meta-priors are enforced by different methods.

In this work, we aim to investigate the use of interventions for studying the characteristics of data representations, such as the aforementioned meta-priors. As a first step towards this goal, we focus on disentanglement and define a novel measure called *Interventional Score*. Our score quantifies not only the usual disentanglement of the representations but also the proportion of the generative factors that are *covered* by them. Preliminary experiments on image data have shown that this is indeed a promising research direction.

2 Interventions in Latent Spaces

In this section, we present the idea behind interventions and demonstrate their usefulness by defining a novel Interventional Score for disentanglement depicted in Fig. 2.

Causal data and Interventions. As real-world data often has an underlying unknown causal structure, we work with datasets for which we can define a Structural Causal Model (SCM). An SCM consists of the set U of exogenous variables representing independent generative factors, a set V of endogenous variables that are descendants of U, and a set of functions f assigning a value to each endogenous variable V given the values of the rest of the variables in the model. An example of a causal dataset is an augmented dSprites [15] dataset which adheres to an underlying causal structure defined in SCM (1). The exogenous random variables X and Y represent the position of an object in an image and have categorical distributions with 8 categories. The random variable O represents the object's orientation which is dependent on its position. The associated causal graph is visualised in Fig. 1.

Causal dSprites SCM (1)

$$U = \{X, Y\}, \qquad V = \{O, I\}, \qquad F = \{f_o, f_I\}$$
$$f_O(X, Y) = \lfloor X \cdot Y \rfloor = O$$
$$f_I(X, Y, O) = I$$

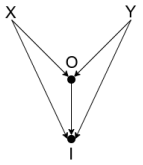

Fig. 1: Causal graph of SCM (1)

Given an SCM, we can now perform *data interventions* which are obtained by holding one generative factor constant and randomly sampling the rest. Similarly as in [22], we consider the set U of exogenous variables as the set of valid interventions [17]. We denote by $\mathscr{A}_D$ the collection of sets containing attributes corresponding to each individual data intervention. These attributes can be arbitrary: images, labels obtained from a classifier or any other features. Our goal is to use interventions on a SCM to study representations of the data. For example, when considering the disentanglement of latent representations we expect each exogenous variable to be encoded in exactly one latent dimension. In this case, we can also perform *latent interventions* by holding one latent dimension constant and randomly sampling the rest from the latent prior distribution. We denote by $\mathscr{A}_Z$ the resulting collection of sets containing attributes corresponding to each individual latent intervention.

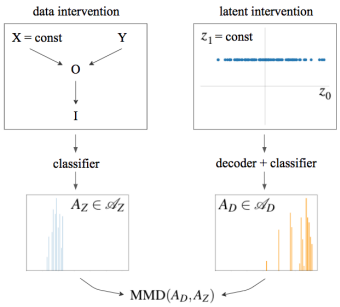

Fig. 2: Interventional score.

Interventional Score. We define a novel measure of disentanglement inspired by the concept of interventions on both the underlying data SCM and latent dimensions. Intuitively, our measure is derived from the following observation: if latent interventions on a dimension produced only limited attributes then this dimension encodes a generative factor. On the other hand, if interventions on a latent dimension produced uniformly distributed attributes, then it does not encode any generative factor. This idea is realised in the Interventional Score by measuring the similarity between every set of attributes $A_D \in \mathscr{A}_D$ obtained from the data interventions and every set of attributes $A_Z \in \mathscr{A}_Z$ obtained from the latent interventions using any test statistics T. For example, T can be Maximum Mean Discrepancy [7] (MMD), any metric between two distribution such as Hilbert Schmidt Independence Criterion [8], any instance of f-divergences [18] or other hypothesis test such as Kolmogorov-Smirnov or Welch's t-test.

More formally, our score is based on calculating $\mathrm{T}(A_D, A_Z)$ for every combination of $A_D \in \mathscr{A}_D$ and $A_Z \in \mathscr{A}_Z$. For each latent *dimension* we can then extract two values: (1) the generative factor whose data interventions yield the lowest T for all latent interventions on the given dimension, and (2) the number of distinct attributes of the generative factor identified in (1) that are covered in the given latent dimension. Therefore, we can not only determine which factor we intervened on but also which attributes of the generative factor this part of the latent space encodes. These values can therefore be though of as *disentangling precision* measuring the usual disentanglement and *disentangling recall* measuring the coverage of the generative factors, respectively. The obtained 2-fold information is summarised for every pair of latent dimension and generative factor in our novel *Interventional Score*.

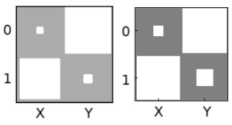

Fig. 3: Hinton diagrams, ours (left) and [5] (right).

Experiment We validate our Interventional Score on a VAE [11], [19] with a 2-dimensional latent space trained on the Causal dSprites dataset. Note that modeling of the causal dependency of O on X and Y is left to the model. We compute the Interventional Score using the unbiased estimator of the squared MMD as given in Lemma 6 in [7] with an exponential kernel. As attributes, we use labels corresponding to the categories of generative factors. For generated images, these are obtained from a classifier trained on the VAE reconstructed images of Causal dSprites. We visualise the Interventional score obtained for each pair of latent dimension and generative factor in Fig. 3 (left) compared with [5] (right).

3 Open Questions

Knowledge of generative factors. The key drawback of the existing methods for analysing representations, especially disentanglement, is that the underlying data generative fac-

tors need to be known a priori. This restricts the current evaluations frameworks to synthetic data for which the generative factors can be manually defined.

Out-of-distribution samples. The current realization of the Interventional Score does not take into account the fact that a latent intervention can result in an attribute that lies outside the training attributes distribution. One way of addressing this is to train a discriminator that can distinguish between the valid and invalid attributes. This could provide additional information about the quality of the attributes, similar to the visual inspection with latent traversal which is commonly used when working with image data.

Extension to other meta-priors. In the future, we wish to investigate how we can leverage interventions in the latent space to obtain sensible measures that also capture other meta-priors. We also plan to investigate how certain information can be retained inside the latent space rather than in the encoder and the decoder which is currently the case when approximating the causal graph.

References

1. Ansari, A. F. & Soh, H. *Hyperprior induced unsupervised disentanglement of latent representations* in *Proceedings of the AAAI Conference on Artificial Intelligence* **33** (2019), 3175–3182.
2. Bengio, Y., Courville, A. & Vincent, P. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence* **35,** 1798–1828 (2013).
3. Burgess, C. P. *et al.* Understanding disentangling in beta-VAE. *arXiv preprint arXiv:1804.03599* (2018).
4. Chen, T. Q., Li, X., Grosse, R. B. & Duvenaud, D. K. *Isolating sources of disentanglement in variational autoencoders* in *Advances in Neural Information Processing Systems* (2018), 2610–2620.
5. Eastwood, C. & Williams, C. K. A framework for the quantitative evaluation of disentangled representations (2018).
6. Esmaeili, B. *et al.* Hierarchical disentangled representations. *stat* **1050,** 12 (2018).
7. Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. & Smola, A. A Kernel Two-Sample Test. *Journal of Machine Learning Research* **13,** 723–773. `http://jmlr.org/papers/v13/gretton12a.html` (2012).
8. Gretton, A., Bousquet, O., Smola, A. & Schölkopf, B. *Measuring statistical dependence with Hilbert-Schmidt norms* in *International conference on algorithmic learning theory* (2005), 63–77.
9. Higgins, I. *et al.* beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. *Iclr* **2,** 6.
10. Kim, H. & Mnih, A. Disentangling by factorising. *arXiv preprint arXiv:1802.05983* (2018).
11. Kingma, D. P. & Welling, M. Auto-encoding variational bayes. *Int. Conf. Learn. Represent.* (2015).
12. Kumar, A., Sattigeri, P. & Balakrishnan, A. Variational inference of disentangled latent concepts from unlabeled observations. *arXiv preprint arXiv:1711.00848* (2017).
13. Lapuschkin, S. *et al.* Unmasking clever hans predictors and assessing what machines really learn. *Nature communications* **10,** 1–8 (2019).
14. Lipton, Z. C. The mythos of model interpretability. *Queue* **16,** 31–57 (2018).
15. Matthey, L., Higgins, I., Hassabis, D. & Lerchner, A. *dSprites: Disentanglement testing Sprites dataset* 2017.
16. Nissim, M., van Noord, R. & van der Goot, R. Fair is better than sensational: Man is to doctor as woman is to doctor. *arXiv preprint arXiv:1905.09866* (2019).
17. Pearl, J. *Causality* (Cambridge university press, 2009).
18. Rényi, A. *et al. On measures of entropy and information* in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics* (1961).
19. Rezende, D. J., Mohamed, S. & Wierstra, D. *Stochastic Backpropagation and Approximate Inference in Deep Generative Models* in *Int. Conf. Mach. Learn.* (2014), 1278–1286.
20. Ridgeway, K. & Mozer, M. C. *Learning deep disentangled embeddings with the f-statistic loss* in *Advances in Neural Information Processing Systems* (2018), 185–194.
21. Samek, W., Wiegand, T. & Müller, K.-R. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296* (2017).
22. Suter, R., Miladinović, D., Schölkopf, B. & Bauer, S. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. *arXiv preprint arXiv:1811.00007* (2018).
23. Thomas, V. *et al.* Independently controllable factors. *arXiv preprint arXiv:1708.01289* (2017).
24. Tschannen, M., Bachem, O. & Lucic, M. Recent advances in autoencoder-based representation learning. *arXiv preprint arXiv:1812.05069* (2018).
25. Xu, J. & Durrett, G. Spherical latent spaces for stable variational autoencoders. *arXiv preprint arXiv:1808.10805.*
26. Zhang, Q., Nian Wu, Y. & Zhu, S.-C. *Interpretable convolutional neural networks* in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), 8827–8836.