A Commentary on the Unsupervised Learning of Disentangled Representations

Francesco Locatello,[2,3] **Stefan Bauer,**[3] **Mario Lucic,**[1] **Gunnar Rätsch,**[2]
Sylvain Gelly,[1] **Bernhard Schölkopf,**[3] **Olivier Bachem**[1]
[1]Google Research, Brain Team
[2]ETH Zurich, Department for Computer Science
[3]Max-Planck Institute for Intelligent Systems
Correspondence to francesco.locatello@inf.ethz.ch and bachem@google.com

Abstract

The goal of the *unsupervised* learning of *disentangled* representations is to separate the independent explanatory factors of variation in the data without access to supervision. In this paper, we summarize the results of (Locatello et al. 2019b) and focus on their implications for practitioners. We discuss the theoretical result showing that the unsupervised learning of disentangled representations is fundamentally impossible without inductive biases and the practical challenges it entails. Finally, we comment on our experimental findings, highlighting the limitations of state-of-the-art approaches and directions for future research.

Introduction

In representation learning, we often have access to high-dimensional observations $\mathbf{x}$ (e.g., images or videos) without additional annotations. However, such observations are often assumed to be the manifestation of a set of low dimensional ground truth factors of variations $\mathbf{z}$. For example, the factors of variation in natural images may be pose, content, location of objects, and lighting conditions.

The goal of representation learning is to learn a vector $r(\mathbf{x})$ which is low dimensional and useful for any downstream task (Bengio, Courville, and Vincent 2013). The key idea of disentangled representations is that they capture the information about the explanatory factors of variations independently: each factor of variation is separately represented in just a few dimensions of the representation (Bengio, Courville, and Vincent 2013). In our example of natural images, we may wish to encode separately pose, content, location of objects, and lighting conditions.

Disentangled representations hold the promise to be both interpretable, robust, and to simplify downstream prediction tasks (Bengio, Courville, and Vincent 2013). Recently, disentanglement has been found useful for a variety of downstream tasks including fair machine learning (Locatello et al. 2019a; Creager et al. 2019), abstract visual reasoning tasks (van Steenkiste et al. 2019) and real-world robotic data sets (Gondal et al. 2019).

State-of-the-art approaches for unsupervised disentanglement learning are largely based on variants of *Variational Autoencoders (VAEs)* (Kingma and Welling 2014) where the encoder is further regularized to encourage disentanglement.

In this paper, we comment on some of the key contributions of (Locatello et al. 2019b):

- We discuss why it is impossible to learn disentangled representations for arbitrary data sets without supervision or inductive biases.
- We provide a sober look at the performances of state-of-the-art approaches. We highlight challenges for model selection and identify critical areas for future research.
- To facilitate future research and reproducibility, we release a library to train and evaluate disentangled representations on standard benchmark data sets.

Background

For the purpose of disentanglement learning, the world is modeled as a two-step generative process. First, we sample a latent variable $\mathbf{z}$ from a distribution with factorized density $p(\mathbf{z}) = \prod_{i=1}^{d} p(\mathrm{z}_i)$. Each dimension of $\mathbf{z}$ corresponds to an independent factor of variation such as pose, content, locations of objects and lighting conditions in an image. Second, the observations are obtained as samples from $p(\mathbf{x}|\mathbf{z})$.

The goal of disentanglement is to encode the factors of variation in a vector $r(\mathbf{x})$ independently. The key idea is that a change in a dimension of $\mathbf{z}$ corresponds to a change in a dimension (or subset of dimensions) of $r(\mathbf{x})$ (Bengio, Courville, and Vincent 2013). This definition has been further extended in the languages of group theory (Higgins et al. 2018) and causality (Suter et al. 2019).

Metrics The lack of a formal definition of disentanglement resulted in a variety of different metrics. We assume access to $\mathbf{z}$ and characterize the structure of the statistical relations between $\mathbf{z}$ and $r(\mathbf{x})$. Intuitively, we measure how the information about $\mathbf{z}$ is encoded in $r(\mathbf{x})$. The *BetaVAE* (Higgins et al. 2017) and *FactorVAE* (Kim and Mnih 2018) scores measures disentanglement by predicting the index of a fixed factor. Other scores are typically composed of two steps: first, they estimate a matrix relating $\mathbf{z}$

and $r(\mathbf{x})$. The *Mutual Information Gap (MIG)* (Chen et al. 2018) and *Modularity* (Ridgeway and Mozer 2018) estimate the pairwise mutual information matrix, *DCI Disentanglement* (Eastwood and Williams 2018) the feature importance predicting $\mathbf{z}$ from $r(\mathbf{x})$ and the *SAPscore* (Kumar, Sattigeri, and Balakrishnan 2018) the predictability of $\mathbf{z}$ from $r(\mathbf{x})$. Second, this matrix is aggregated to obtain a score by computing some normalized gap either row- or column-wise. For more details, see Appendix C of (Locatello et al. 2019b).

Methods In *Variational Autoencoders (VAEs)* (Kingma and Welling 2014), one assumes a prior $P(\mathbf{z})$ on the latent space and parameterizes the conditional probability $P(\mathbf{x}|\mathbf{z})$ using a deep neural network (i.e., a *decoder network*). The posterior distribution is approximated by a variational distribution $Q(\mathbf{z}|\mathbf{x})$, again parameterized using a deep neural network (i.e., an *encoder network*). The model is then trained by maximizing a variational lower-bound to the log-likelihood and the representation $r(\mathbf{x})$ is usually taken to be the mean of the encoder distribution. To learn disentangled representations, state-of-the-art approaches enrich the VAE objectives with a suitable regularizer.

The β-VAE (Higgins et al. 2017) and Annealed-VAE (Burgess et al. 2018) constrain the capacity of the VAE bottleneck. The intuition is that recovering the factors of variation is the most efficient compression scheme to achieve good reconstruction (Peters, Janzing, and Schölkopf 2017). The Factor-VAE (Kim and Mnih 2018) and β-TCVAE both penalize the total correlation of the aggregated posterior $Q(\mathbf{z}) = \int Q(\mathbf{z}|\mathbf{x})d\mathbf{x}$ (i.e. the encoder distribution after marginalizing the training data). The DIP-VAE variants (Kumar, Sattigeri, and Balakrishnan 2018) match the moments of the aggregated posterior and a "disentanglement prior", which in practice is simply a factorized distribution. We refer to Appendix B of (Locatello et al. 2019b) for a more detailed description.

Theoretical impossibility

Theorem 1 in (Locatello et al. 2019b) states that the unsupervised learning of disentangled representation is impossible for arbitrary data sets. Even in the infinite data regime, where supervised learning algorithms such as k-nearest neighbours classifiers are consistent, no model can find a disentangled representation observing samples from $P(\mathbf{x})$ only. This theoretical result motivates the need for either implicit supervision, explicit supervision, or inductive biases.

The key idea is that we can construct two generative models whose latent variables $\mathbf{z}$ and $f(\mathbf{z})$ are entangled with each other but have the same marginal distribution over $\mathbf{x}$, i.e., the same $P(\mathbf{x})$. If a representation is disentangled with one of these generative models it must be entangled with the other by construction. Observing only samples from $P(\mathbf{x})$, it is impossible to distinguish which model $r(\mathbf{x})$ should disentangle: both $\mathbf{z}$ and $f(\mathbf{z})$ are equally plausible and "look the same" as they produce the same $\mathbf{x}$ with the same probability.

Note that Theorem 1 in (Locatello et al. 2019b) does not account for the structure that real world generative models may exhibit. Inductive biases on both the models and the data may be sufficient to learn disentangled representations in practice as certain solutions may be favored instead of others, i.e., some model may naturally converge to a solution that disentangles the true $\mathbf{z}$ instead of $f(\mathbf{z})$. Similar results have been obtained in the context of nonlinear ICA (Hyvärinen and Pajunen 1999) where i.i.d. data is known to be insufficient for identifiability, in general.

Implications We proved that the unsupervised learning of disentangled representations is in general impossible without inductive biases on both methods and data sets. We argue that future work should make the role of inductive biases or supervision more explicit.

Disentanglement in practice

In this section, we highlight the implications of some of the empirical results of (Locatello et al. 2019b). We implemented six recent unsupervised disentanglement learning methods as well as six disentanglement metrics from scratch. Overall, we trained over 12000 models and computed over 150000 scores on seven data sets and 50 random seeds.[1] We refer to Section 5 of (Locatello et al. 2019b) for more details and a richer quantitative description.

Which method should be used?

This first question is particularly relevant for practitioners interested in the benefits of disentanglement methods off-the-shelf. In Figure 1 (left), e observe that the choice of the objective function seems to matter less than the choice of hyperparameters and seed. In particular, only 37% of the observed variance in the models can be explained by the choice of the objective function. Since our trained models exhibit such a large variance, it appears to be crucial to identify good hyperparameters and runs.

Implications It is not clear which method should be used and choosing good hyperparameters, and selecting good runs seem to be matter more.

How to choose the hyperparameters?

We investigated whether we may find "rules of thumb" for selecting good hyperparameters. In Figure 1 (center), we plot the median FactorVAE score for different regularization strengths for each method on Cars3D. We observe that no method is consistently better than all the others and there does not seem to be an obvious trend that can be used to maximize disentanglement scores. In Figure 1 (right), we test whether good hyperparameter settings may be transferred across data sets. We observe that at the distribution level there appears to be some correlation between the disentanglement scores across the different data sets.

Implications There is no clear rule of thumb, but transfer across data sets may help. Note that we still cannot distinguish between a good and a bad training run.

[1] Reproducing our results requires approximately 2.52 GPU years (NVIDIA P100).

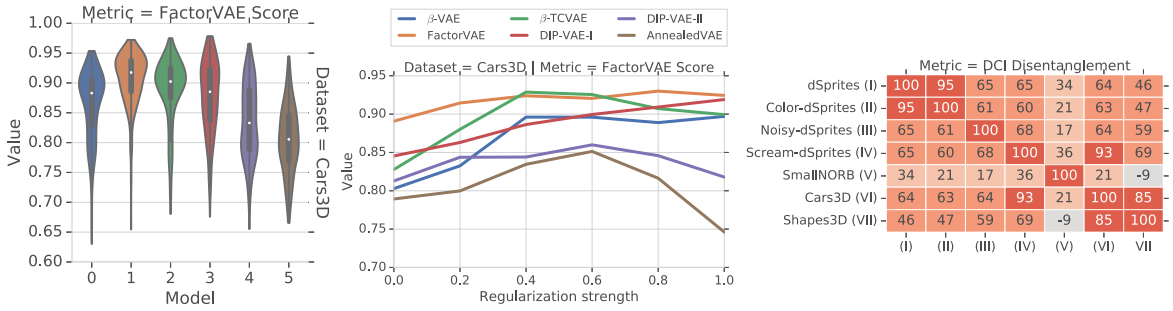

Figure 1: (left) FactorVAE score for each method on Cars3D. Models are abbreviated: 0=β-VAE, 1=FactorVAE, 2=β-TCVAE, 3=DIP-VAE-I, 4=DIP-VAE-II, 5=AnnealedVAE. The variance is due to different hyperparameters and random seeds. We observe that the scores are heavily overlapping. (center) FactorVAE score vs hyperparameters for each score on Cars3d. There seems to be no model dominating all the others and for each model there does not seem to be a consistent strategy in choosing the regularization strength. (right) Rank-correlation of DCI disentanglement metric across different data sets. Good hyperparameters seem to transfer especially between dSprites and Color-dSprites.

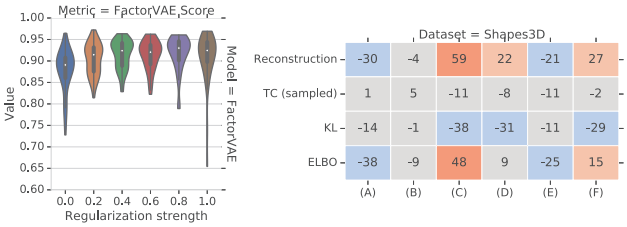

Figure 2: (left) Distribution of FactorVAE scores for FactorVAE model for different regularization strengths on Cars3D. (right) Rank correlation between unsupervised scores and disentanglement metrics on Shapes3D. Metrics are abbreviated: (A)=BetaVAE Score, (B)=FactorVAE Score, (C)=MIG , (D)=DCI Disentanglement, (E)=Modularity, (F)=SAP.

How to select the best model from a set of trained models?

First, we note that the transfer of hyperparmeters does not reliably outperforms random model selection: it improves only 59.3% of the times. To understand why this is the case we plot in Figure 2 (left) the distribution of Factor VAE models evaluated with the FactorVAE score on Cars3D. We observe that randomness has a substantial impact on the representation as a good run with bad hyperparameters can easily outperform a bad run with the best hyperparameters. Finally, we check whether the unsupervised training metrics may be used for model selection. In Figure 2 (right), we observe that the training metrics appear to be rather uncorrelated with disentanglement.

Implications Unsupervised model selection remains an open research challenge. Transfer of good hyperparameters does not seem to work and we did not find a way to distinguish between good and bad runs without supervision.

Directions of future research

Finally, we discuss the critical open challenges in disentanglement and some of the lessons we learned with this study.

Inductive biases and implicit and explicit supervision. Our results highlights an overall need for supervision. In theory, inductive biases are crucial to distinguish among equally plausible generative models. In practice we did not find a reliable strategy to choose hyperparameters without supervision. Recent work (Duan et al. 2019) proposed a stability based heuristic for unsupervised model selection. Further exploring these techniques may help us understand the practical role of inductive biases and implicit supervision. Otherwise, we advocate to consider different settings, for example when limited explicit (Locatello et al. 2019c) or weak supervision (Bouchacourt, Tomioka, and Nowozin 2018; Gresele et al. 2019) is available.

Experimental setup and diversity of data sets. Our study highlights the need for a sound, robust, and reproducible experimental setup on a diverse set of data sets. In our experiments, we observed that the results may be easily misinterpreted if one only looks at a subset of the data sets. As current research is typically focused on the synthetic data sets of (Higgins et al. 2017; Reed et al. 2015; LeCun, Huang, and Bottou 2004; Kim and Mnih 2018; Locatello et al. 2019b) — with only a few recent exceptions (Gondal et al. 2019) — we advocate for insights that generalize across data sets rather than individual absolute performance. For this reason, we released `disentanglement_lib`[2], a library to facilitate repro-

[2]https://github.com/google-research/disentanglement_lib

ducible research on disentanglement. Our library allows to train and evaluate state-of-the-art disentangled representations on common benchmark data sets and produces automatic visualizations for visual inspection on all the trained models. Furthermore, we released over 10000 trained models, which can be used as baselines for future research.

Acknowledgements

The authors thank Ilya Tolstikhin, Paul Rubenstein and Josip Djolonga for helpful discussions and comments. This research was partially supported by the Max Planck ETH Center for Learning Systems, by an ETH core grant (to Gunnar Rätsch) and a Google Ph.D. Fellowship to FL. This work was partially done while FL was at Google Research Zurich.

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.

Bouchacourt, D.; Tomioka, R.; and Nowozin, S. 2018. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. In *AAAI Conference on Artificial Intelligence*.

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

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

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.

Duan, S.; Watters, N.; Matthey, L.; Burgess, C. P.; Lerchner, A.; and Higgins, I. 2019. A heuristic for unsupervised model selection for variational disentangled representation learning. *arXiv preprint arXiv:1905.12614*.

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

Gondal, M. W.; Wüthrich, M.; Miladinović, 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 *Advances in Neural Information Processing Systems*.

Gresele, L.; Rubenstein, P. K.; Mehrjou, A.; Locatello, F.; and Schölkopf, B. 2019. The incomplete rosetta stone problem: Identifiability results for multi-view nonlinear ica. In *Conference on Uncertainty in Artificial Intelligence (UAI)*.

Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; and Lerchner, A. 2017. beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*.

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

Hyvärinen, A., and Pajunen, P. 1999. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*.

Kim, H., and Mnih, A. 2018. Disentangling by factorising. In *International Conference on Machine Learning*.

Kingma, D. P., and Welling, M. 2014. Auto-encoding variational Bayes. In *International Conference on Learning Representations*.

Kumar, A.; Sattigeri, P.; and Balakrishnan, A. 2018. Variational inference of disentangled latent concepts from unlabeled observations. In *International Conference on Learning Representations*.

LeCun, Y.; Huang, F. J.; and Bottou, L. 2004. Learning methods for generic object recognition with invariance to pose and lighting. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Locatello, F.; Abbati, G.; Rainforth, T.; Bauer, S.; Schölkopf, B.; and Bachem, O. 2019a. On the fairness of disentangled representations. In *Advances in Neural Information Processing Systems*.

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

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

Peters, J.; Janzing, D.; and Schölkopf, B. 2017. *Elements of Causal Inference - Foundations and Learning Algorithms*. Adaptive Computation and Machine Learning Series. MIT Press.

Reed, S.; Zhang, Y.; Zhang, Y.; and Lee, H. 2015. Deep visual analogy-making. In *Advances in Neural Information Processing Systems*.

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

Suter, R.; Miladinović, D.; Bauer, S.; and Schölkopf, B. 2019. Interventional robustness of deep latent variable models. In *International Conference on Machine Learning*.

van Steenkiste, S.; Locatello, F.; Schmidhuber, J.; and Bachem, O. 2019. Are disentangled representations helpful for abstract visual reasoning? In *Advances in Neural Information Processing Systems*.