ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

Modular Representations for Weak Disentanglement

Andrea Valenti and Davide Bacciu *

University of Pisa - Department of Computer Science
Largo B. Pontecorvo, 3 56127 Pisa - Italy

Abstract. The recently introduced weakly disentangled representations proposed to relax some constraints of the previous definitions of disentanglement, in exchange for more flexibility. However, at the moment, weak disentanglement can only be achieved by increasing the amount of supervision as the number of factors of variations of the data increase. In this paper, we introduce modular representations for weak disentanglement, a novel method that allows to keep the amount of supervised information constant with respect the number of generative factors. The experiments shows that models using modular representations can increase their performance with respect to previous work without the need of additional supervision.

1 Introduction

Intuitively speaking, a *disentangled representation* can be defined as a (usually low-dimensional) encoding of z of a data sample x, where distinct components of z are responsible for encoding a specific generative factor of the data. Despite the different attempts in the literature, coming up with a formal definition of what disentanglement actually is has proven more difficult than expected [1]. Several works just assume that a disentangled representation is a representation in which a *single* latent dimension responsible for encoding a *single* generative factor of the data. This definition, while easy to formalise in a mathematical way, has resulted to be too restrictive in general. Recently, [2] relaxed this definition by introducing *weak disentangled* representation, where each generative factor can be encoded in a different region of the latent space without imposing additional limitations on their dimensionality. Despite the advantages of this new approach, the initial implementation of [2] suffered from the fact that the number of annotations required for achieving weak disentanglement grew very quickly in the number of generative factors.

In this paper, we address this limitation by introducing *modular representations* for weak disentanglement. In a modular representation, each partition of the latent space encodes the respective generative factor in a different adaptive prior distribution, independent from the others. We show that models that use modular representations are able to accurately perform controlled manipulations of the learned generative factors of the data without the need of increasing the amount of supervised information.

*The work has been partially supported by the EU H2020 TAILOR project (n.952215).

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

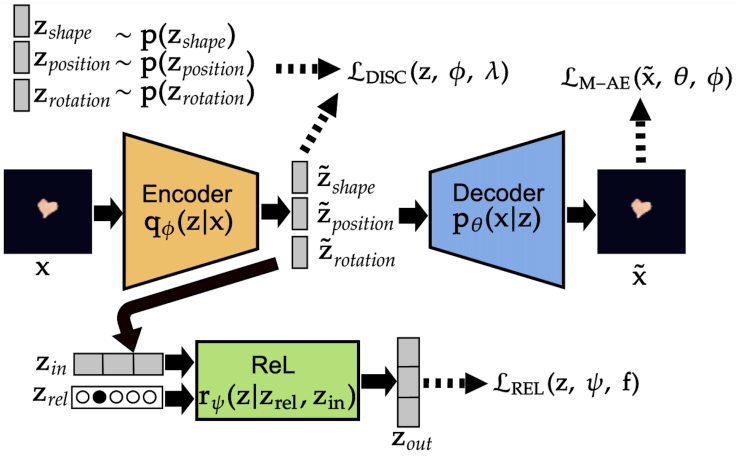

Fig. 1: Overview of the model's architecture.

2 Related Works

Early methods for disentanglement are mainly concerned with increasing prior regularisation of the loss function [3, 4]. Another line of work [5, 6, 7] penalises different terms of the same loss function in various ways. They define disentanglement using simple mathematical notions (e.g. total correlation of the latent dimensions). After the results of [8], showing that pure unsupervised disentanglement is in general impossible to achieve, many works started using various degrees of supervised information [9], either in the form of complete supervision on a small subset of training data [10] or partial annotations on a subset of generative factors [11]. Some works use additional classifier networks on the latent space in order to separate different parts of the latent codes. While useful, these method are not practical when multiple factors of variations need to be disentangled at the same time. Other methods for introducing implicit supervision involve dropping the i.i.d. assumption by leveraging relational information between the samples. The relational information can be group-wise [12], pair-wise [13], or structural [14]. Recently, [2] introduced the concept of weak disentanglement, overcoming many of the above limitations. However, their method requires an increasing amount of supervision when the number of generative factors increases.

3 Modular Representations for Weak Disentanglement

A general overview of the model's architecture is illustrated in Fig. 1. We frame our representation learning problem as an auto-encoding task. Given a data sample $x \sim p(x)$, we want to output a faithful reconstruction $\hat{x}$. The *encoder* network $q_\phi(z|x)$, parameterised by ϕ, takes a data sample x as input and produces G latent codes, where G is the number of generative factors of

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

the data. Conversely, the *decoder* network $p_\theta(x|z_1, z_2, ..., z_G)$, parameterised by θ, combines these partial latent codes to reconstruct the initial input. The resulting *Modular AutoEncoder* (M-AE) model is then trained using the following maximum likelihood objective:

$$\max_{\theta,\phi} \mathcal{L}_{M-AE}(x, \theta, \phi) = \mathbb{E}_{q_\phi(z|x)} \left[\log p_\theta(x|z)\right] - \beta \sum_{g_i=1}^{G} \mathrm{KL}\left(q_\phi(z_{g_i})||p(z_{g_i})\right). \quad (1)$$

The first term of Eq.1 is directly responsible for ensuring a good reconstruction of the original input x. The second term, a sum of KL divergences between the aggregate posteriors $q_\phi(z_i)$ and the priors $p(z_i)$, encourages each partition of the latent space z_i to follow a specific prior distribution. These priors are directly inspired from the data and are enforced in an adversarial way, similar to GANs [15]. In particular, this term is optimised via an additional discriminator network $d_\psi(z)$, parameterised by λ:

$$\min_{\phi} \max_{\lambda} \quad \mathcal{L}_{DISC}(z, \phi, \lambda) = \mathbb{E}_{q_\phi(z)}[\log d_\lambda(z)] + \mathbb{E}_{p(z)}[\log(1 - d_\lambda(z))]. \quad (2)$$

This adversarial loss allows us to choose the most suitable prior distribution for each partition. In particular, since our goal is to identify all possible real-world instances of a particular value of a generative factor g_i, we model each $p(z_i)$ as a *mixture of normal distributions*: $p(z_{g_i}) = \frac{1}{V_{g_i}} \sum_{v=1}^{V_{g_i}} \mathcal{N}(\mu_{g_i,v}, \Sigma^2_{g_i,v})$, where V_{g_i} is the number of values that factor g_i can take. We build a different mixture for each partition. Specifically, the parameters μ and Σ^2 of each prior's components are empirically estimated using a small subset of annotated samples: $\forall i, v. \mu_{i,v} = \mathbb{E}\left[\{z\}_{g_i=v}\right], \quad \Sigma^2_{i,v} = \mathrm{Var}\left[\{z\}_{g_i=v}\right]$, where $\{z\}_{g_i=v}$ denotes the subset of (encoded) supervised samples where the factor g_i takes value v. Since each latent partition encodes a different generative factor, when can re-use the same annotated samples for computing the different parts of the prior. The second part of the model is the *Relational Learner* (ReL). During training, the ReL learns how to perform controlled changes to specific properties of the data sample by leveraging the representations learned by the M-AE. The ReL is composed of the *relational sub-network* $r_\psi(z|z_{rel}, z_{in})$, parameterised by ψ. Assuming that the relation to be learned affects only the value of a single factor of variations, the relational objective becomes the following:

$$\max_{\psi} \quad \mathcal{L}_{ReL}(z, f, \psi) = \log p\left(z_{f(g_i)}\right) \sum_{j \neq i} \log p\left(z_{g_j}\right) \quad (3)$$

where $z = [z_{g_1} \ldots, z_{g_i}, \ldots z_{g_G}] \sim r_\psi(z|z_{rel}, z_{in})$ is the output of the relational learner. The function f defines the "connections" between the prior components that correspond to a specific relation. This can be easily extended for losses that affect multiple factors. This loss function encourages the partition affected by the relation to match the prior of the new value of that factor, while the other partitions remain unchanged. The correspondence between components of the prior and generative factor values is made possible by the representation learned

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

by the M-AE. Finally, training is done end-to-end by combining the previous losses $\mathcal{L} = \mathcal{L}_{M-AE}(x, \theta, \phi, \lambda) + \mathcal{L}_{ReL}(z, \psi, f)$,

4 Experiments[1]

Datasets. We consider two disentanglement tasks based on the dSprites [16] and Shapes3D [17] datasets, containing respectively 2D and 3D images of shapes that express different combinations of generative factors (`shape`, `x/y-position`, `scale`, and `orientation` for dSprites; `floor-color`, `shape-color`, `background -color`, and `orientation` for Shapes3D). We consider all the relations that affect the change of a single generative factor of the data (e.g. `move-left, move-right, +hue, change-shape, etc.`). No restriction is imposed on the nuisance factors, that are able to vary freely when applying relations on the latent codes. For each dataset, we construct three versions of increasing complexity, characterised by different choices of relevant and nuisance factors.

Training Setting. The M-AE encoder and decoder are implemented as a CNN, while the prior's Discriminator and the ReL are 3-layers MLP with 1024 units each. We use 8-dimensional latent codes for each generative factors, for a maximum size of the latent space of $N_z = 32$. All tasks use a batch size of 1024 for the M-AE and 128 for the ReL. The parameter β of Eq. 1 is set to 0.1. The optimiser used for all modules is Adam with a learning rate of 10^{-4}. Training is divided in two stages. In the first stage, called *warmup*, only the M-AE is trained. The prior is set to $\forall i. p(z_i) \sim \text{Uniform}(-1, 1)$. After 1000 epochs we enter the *full training* stage, where the prior of each latent partition is set to the adaptive prior described in Sec. 3. We construct a different prior for each generative factor, leveraging the annotations of the supervised subset. At the same time, the training of the ReL begins: the input data samples are constructed as triples $(z_{in}, z_{rel}, z_{out})$, where z_{in} and z_{out} are respectively the encoded input and output samples for the relation z_{rel}. The latent codes are sampled from their respective components in the latent space. The concurrent training of the M-AE and the ReL is carried on during the *full training* phase for 5000 additional epochs.

Latent Codes Manipulation. In this first set of experiments, we are interested to analyse how well suited are the modular representations to perform controlled changes of generative factors in the latent codes. We compute the relation accuracy of the ReL by first sampling a latent code from the prior, then we apply a random relation and check the outcome. The results are reported in Table. 1, compared with the previous work of [2]. The results show that modular representations are beneficial for the accuracy of the ReL, while not requiring an increasing amount of supervised data when the number of factor value combinations increases.

[1] All the code of the models and the experiments is publicly available: `https://github.com/Andrea-V/Weak-Disentanglement` .

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

Table 1: Relational accuracy of the ReL. τ is the number of supervised samples used during training.

		Previous Work [2]		This Work	
	Factor Combinations	Accuracy	τ	Accuracy	τ
dSprites.v2	27	0.592	270	**0.724**	1000
dSprites.v3	135	0.571	1350	**0.751**	1000
dSprites.v4	1080	0.491	10800	**0.685**	1000
Shapes3D.v2	40	0.275	400	**0.690**	1000
Shapes3D.v3	120	0.220	1200	**0.653**	1000
Shapes3D.v5	12000	0.124	12000	**0.633**	1000

Table 2: Disentanglement scores of latent representations. Higher is better.

	dSprites			Shapes3D		
	DCI	MIG	SAP	DCI	MIG	SAP
Locatello et al. [10]	0.533	0.01	0.01	0.48	0.05	0.08
Gabbay et al. [11]	0.8366	0.14	0.57	**1.0**	0.3	**1.0**
Valenti et al. [2]	0.9543	**0.994**	0.7728	0.6921	0.6897	0.5007
Ours	**0.9732**	0.9721	**0.7877**	0.7056	**0.6919**	0.5511

Disentanglement Scores. We compare the SAP [18], DCI [6] and MIG [19] disentanglement scores against several models of the literature. Following the approach of [2], we convert our modular representations into its corresponding generative factor values before computing the scores. This step can be done at no additional computational cost. The results are reported in Table 2 showing that modular representations have a beneficial impact to all the scores, especially considering the challenging SAP score. This is a strong sign that the modular separation of weakly disentangled representations is indeed able to improve the disentanglement performance of generative models.

5 Conclusion

In this paper, we introduced a novel framework for learning *modular weakly disentangled representations.* Modular representations encode each generative factor into a separate partition of the latent space, thus overcoming the need of requiring additional supervision when the number of value combinations of the generative factors increases. The experiments show that modular representations allow to perform controlled manipulations to selected generative factors with high accuracy. This, in turn, results in high disentanglement scores. In the future, we wish to further enhance the expressivity of our methods by finding ways to encode continuous generative factors in a weakly disentangled way.

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

References

[1] Kien Do and Truyen Tran. Theory and evaluation metrics for learning disentangled representations. *arXiv preprint arXiv:1908.09961*, 2019.

[2] Andrea Valenti and Davide Bacciu. Leveraging relational information for learning weakly disentangled representations. *Accepted at IEEE WCCI 2022*, 2022.

[3] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR2016*, 2016.

[4] Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-vae. *2017 NIPS Workshop on Learning Disentangled Representations*, 2017.

[5] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. *Learning Disentangled Representations 2017 NIPS Workshop*, 2017.

[6] Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, pages 2610–2620, 2018.

[7] Shengjia Zhao, Jiaming Song, and Stefano Ermon. Infovae: Balancing learning and inference in variational autoencoders. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 5885–5892, 2019.

[8] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schoelkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of 36th ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 4114–4124. PMLR, 2019.

[9] Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, and Marc'Aurelio Ranzato. Fader networks: Manipulating images by sliding attributes. In *Advances in neural information processing systems*, pages 5967–5976, 2017.

[10] Francesco Locatello, Michael Tschannen, Stefan Bauer, Gunnar Rätsch, Bernhard Schölkopf, and Olivier Bachem. Disentangling factors of variation using few labels. *arXiv preprint arXiv:1905.01258*, 2019.

[11] Aviv Gabbay, Niv Cohen, and Yedid Hoshen. An image is worth more than a thousand words: Towards disentanglement in the wild. *arXiv preprint arXiv:2106.15610*, 2021.

[12] Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *International Conference on Machine Learning*, pages 6348–6359. PMLR, 2020.

[13] Junxiang Chen and Kayhan Batmanghelich. Weakly supervised disentanglement by pairwise similarities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3495–3502, 2020.

[14] Junwen Bai, Weiran Wang, and Carla Gomes. Contrastively disentangled sequential variational autoencoder. *NeurIPS2021*, 2021.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[16] Loic Matthey, Irina Higgins, Demis Hassabis, and Alexander Lerchner. dsprites: Disentanglement testing sprites dataset. https://github.com/deepmind/dsprites-dataset/, 2017.

[17] Chris Burgess and Hyunjik Kim. 3d shapes dataset. https://github.com/deepmind/3dshapes-dataset/, 2018.

[18] Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. *ICLR 2018*, 2017.

[19] Cian Eastwood and Christopher KI Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018.