Disentanglement Challenge: From Regularization to Reconstruction

Jie Qiao* QIAOJIE.CHN@GMAIL.COM
Zijian Li* LEIZIGIN@GMAIL.COM
Boyan Xu* HPAKYIM@GMAIL.COM
Ruichu Cai CAIRUICHU@GDUT.EDU.CN
School of Computers, Guangdong University of Technology, China

Kun Zhang KUNZ1@CMU.EDU
Department of philosophy, Carnegie Mellon University

Abstract

The challenge of learning disentangled representation has recently attracted much attention and boils down to a competition[1]. Various methods based on variational auto-encoder have been proposed to solve this problem, by enforcing the independence between the representation and modifying the regularization term in the variational lower bound. However recent work by Locatello et al. (2018) has demonstrated that the proposed methods are heavily influenced by randomness and the choice of the hyper-parameter. This work is built upon the same framework in Stage 1 (Li et al., 2019), but with different settings; to make it self-contained, we provide this manuscript, which is unavoidably very similar to the report for Stage 1. In detail, in this work, instead of designing a new regularization term, we adopt the FactorVAE but improve the reconstruction performance and increase the capacity of network and the training step. The strategy turns out to be very effective in achieving disentanglement.

Keywords: Disentangled representation, unsupervised learning

1. Introduction

The great success of unsupervised learning heavily depends on the representation of the feature in the real-world. It is widely believed that the real-world data is generated by a few explanatory factors which are distributed, invariant, and disentangled (Bengio et al., 2013). The challenge of learning disentangled representation boils down into a competition[1] to build the best disentangled model.

The key idea in disentangled representation is that the perfect representation should be a one-to-one mapping to the ground truth disentangled factor. Thus, if one factor changed and other factors fixed, then the representation of the fixed factor should be fixed accordingly, while others' representation changed. As a result, it is essential to find representations that (i) are independent of each other, and (ii) align to the ground truth factor.

Recent line of works in disentanglement representation learning are commonly focused on enforcing the independence of the representation by modifying the regulation term in the

* equal contribution
1. https://www.aicrowd.com/challenges/neurips-2019-disentanglement-challenge

© 2019 J. Qiao, Z. Li, B. Xu, R. Cai & K. Zhang.

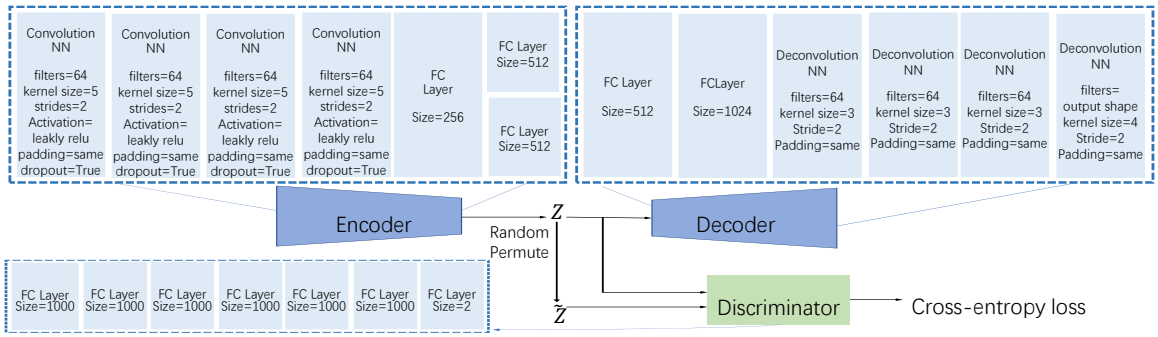

Figure 1: The architecture of FactorVAE, in which FC layer denote the full connection layer

variational lower bound of Variational Autoencoders (VAE) (Kingma and Welling, 2013), including β-VAE (Higgins et al., 2017), AnnealedVAE (Burgess et al., 2018), β-TCVAE (Chen et al., 2018), DIP-VAE (Kumar et al., 2018) and FactorVAE (Kim and Mnih, 2018). See Appendix A for more details of these model.

To evaluate the performance of disentanglement, several metrics have been proposed, including the FactorVAE metric (Kim and Mnih, 2018), Mutual Information Gap (MIG) (Chen et al., 2018), DCI metric (Eastwood and Williams, 2018), IRS metric (Suter et al., 2019), and SAP score (Kumar et al., 2018).

However, one of our findings is that these methods are heavily influenced by randomness and the choice of the hyper-parameter. This phenomenon was also discovered by Locatello et al. (2018). Therefore, rather than designing a new regularization term, we simply use FactorVAE but at the same time improve the reconstruction performance. We believe that, the better the reconstruction, the better the alignment of the ground-truth factors. Therefore, the more capacity of the encoder and decoder network, the better the result would be. Furthermore, after increasing the capacity, we also try to increase the training step which also shows a significant improvement of evaluation metrics. The final architecture of FactorVAE is given in Figure 1. Note that, this work is built upon the same framework in stage 1 (Li et al., 2019), but with different settings; to make it self-contained, we provide this manuscript, which is unavoidably very similar to the report for Stage 1.

Overall, our contribution can be summarized as follow: (1) we found that the performance of the reconstruction is also essential for learning disentangled representation, and (2) we achieve state-of-the-art performance in the competition.

2. Experiments Design

In this section, we explore the effectiveness of different disentanglement learning models and the performance of the reconstruction for disentangle learning. We first employ different kinds of variational autoencoder including *BottleneckVAE*, *AnneledVAE*, *DIPVAE*, *BetaTCVAE*, and *BetaVAE* with 30000 training step. Second, we want to know whether the capacity plays an important role in disentanglement. The hypothesis is that the larger the capacity, the better reconstruction can be obtained, which further reinforces the disentanglement. In detail, we control the number of latent variables.

Table 1: Variational autoencoder with 30000 training steps in Stage 1.

VAE variation	FactorVAE	sap score	dci	irs	mig
BottleneckVAE	0.453	0.0395	0.107	0.547	0.0589
AnneledVAE	0.3586	0.0069	0.1153	0.5122	0.0237
DIPVAE	0.265	0.005	0.021	0.265	0.490
BetaTCVAE	0.342	0.026	0.093	0.342	0.981
BetaVAE	0.3586	0.0069	0.1153	0.5122	0.0237
FactorVAE	**0.449**	**0.0596**	**0.1385**	**0.5976**	**0.0589**

Table 2: FactorVAE with different training step in Stage 1.

training step	FactorVAE	sap score	dci	irs	mig
FactorVAE (30k)	0.449	0.0596	0.1385	0.5976	0.0589
FactorVAE (500k)	0.432	0.0743	0.4395	0.6041	0.0739
FactorVAE (1000k)	**0.4844**	**0.155**	**0.523**	**0.6205**	**0.3887**

3. Experiments Results

In this section, we present our experiment result in stage 1 and stage 2 of the competition. We first present the performance of different kinds of VAEs in stage 1, which is given in Table 1. It shows that FactorVAE achieves the best result when the training step is 30000. In the following experiment, we choose FactorVAE as the base model.

Then, as shown in Table 2, we increase the step size and we find that the best result was achieved at 1000k training steps. The experiment in this part may not be sufficient, but it still suggests the fact that the larger the capacity is, the better the disentanglement performance. Since we increase the capacity of the model, it is reasonable to also increase the training steps at the same time. Furthermore, as shown in Table 3, using sufficient large training step ($\geq 800k$), we investigate the effectiveness of the number of latent variables. This experiment is performed in stage 2 and suggests that the FactorVAE and the DCI metric are positive as the latent variables increase, while the other metrics decrease. The best result in the ranking is marked as bold, which suggests that we should choose an appropriate number of latent variables.

Table 3: Sensitivity analysis for the number of latent variables in Stage 2

Num. of Latent	FactorVAE	sap score	dci	irs	mig
256	0.4458	0.1748	0.5785	0.5553	0.4166
512	**0.5018**	**0.1545**	**0.5457**	**0.6824**	**0.3739**
768	0.4526	0.0942	0.5154	0.4638	0.2793
1024	0.4708	0.0955	0.542	0.4867	0.2781
1536	0.5202	0.008	0.5932	0.5071	0.0058
2048	0.5374	0.0025	0.6351	0.5218	0.0062
3072	0.5426	0.0053	0.6677	0.5067	0.0117

4. Conclusion

In this work, we conducted an empirical study on disentangled learning. We first conduct several experiments with different disentangle learning methods and select the FactorVAE as the base model; and second we improve the performance of the reconstruction, by increasing the capacity of the model and the training step. Finally, our results appear to be competitive.

References

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in beta-vae. *arXiv preprint arXiv:1804.03599*, 2018.

Tian Qi 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.

Cian Eastwood and Christopher KI Williams. A framework for the quantitative evaluation of disentangled representations. 2018.

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. *ICLR*, 2(5):6, 2017.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pages 2654–2663, 2018.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational inference of disentangled latent concepts from unlabeled observations. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H1kG7GZAW.

Zijian Li, Jie Qiao, Boyan Xu, Ruichu Cai, and Kun Zhang. Disentanglement challenge: From regularization to reconstruction. *NeurIPS 2019 Workshop Disentanglement Challenge (Stage 1)*, 2019. URL https://openreview.net/forum?id=SJgy_NNmwB.

Francesco Locatello, Stefan Bauer, Mario Lucic, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. *arXiv preprint arXiv:1811.12359*, 2018.

Raphael Suter, Djordje Miladinovic, Bernhard Schölkopf, and Stefan Bauer. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness. In *International Conference on Machine Learning*, pages 6056–6065, 2019.

Appendix A. Related works

In this section, we are going to summarize the state-of-the-art unsupervised disentanglement learning methods. Most of works are developed based on the Variational Auto-encoder

(VAE) (Kingma and Welling, 2013), a generative model that maximize the following evidence lower bound to approximate the intractable distribution $p_\theta(\mathbf{x}|\mathbf{z})$ using $q_\phi(\mathbf{z}|\mathbf{x})$,

$$\max_{\phi,\theta} \quad \underbrace{\mathbb{E}_{p(\mathbf{x})}[\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})]}_{Reconstruction\ Loss} - \underbrace{D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z}))]}_{Regularization}, \tag{1}$$

where $q_\phi(\mathbf{z}|\mathbf{x})$ denote *Encoder* with parameter ϕ and $p_\theta(\mathbf{x}|\mathbf{z})$ denote *Decoder* with parameter θ.

As shown in Table 4, all the lower bound of variant VAEs can be described as *Reconstruction Loss+ Regularization* where all the Regularization term and the hyper-parameters are given in this table.

Table 4: Summary of variant unsupervised disentanglement learning methods

Model	Regularization	Hyper-Parameters
β-VAE	$\beta\ D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}\|\mathbf{x})\|p(\mathbf{z})\right)]$	β
AnnealedVAE	$\gamma\left\|D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}\|\mathbf{x})\|p(\mathbf{z})\right) - C\right\|]$	γ, C
FactorVAE	$D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}\|\mathbf{x})\|p(\mathbf{z})\right)] + \gamma D_{\mathrm{KL}}\left(q(\mathbf{z})\|\prod_{j=1}^{d} q\left(\mathbf{z}_j\right)\right)$	γ
DIP-VAE-I	$D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}\|\mathbf{x})\|p(\mathbf{z})\right)] + \lambda_{od}\sum_{i\neq j}\left[\mathrm{Cov}_{p(\mathbf{x})}\left[\mu_\phi(\mathbf{x})\right]\right]_{ij}^2 + \lambda_d\sum_i\left(\left[\mathrm{Cov}_{p(\mathbf{x})}\left[\mu_\phi(\mathbf{x})\right]\right]_{ii} - 1\right)^2$	λ_{od}, λ_d
DIP-VAE-II	$D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}\|\mathbf{x})\|p(\mathbf{z})\right)] + \lambda_{od}\sum_{i\neq j}\left[\mathrm{Cov}_{q_\phi}[\mathbf{z}]\right]_{ij}^2 + \lambda_d\sum_i\left(\left[\mathrm{Cov}_{q_\phi}[\mathbf{z}]\right]_{ii} - 1\right)^2$	λ_{od}, λ_d