The 34th Annual Conference of the Japanese Society for Artificial Intelligence, 2020

# Sparsity enforcement on latent variables for better disentanglement in VAE

Paulino Crsitovao [*1*2] Hidemoto Nakada [*2*1] Yusuke Tanimura [*2*1] Hideki Asoh [*2]

[*1] University of Tsukuba
[*2] National Institute of Advanced Industrial Science and Technology of Japan

We address the problem of unsupervised latent factorization and reconstruction accuracy. The related work on unsupervised representations focuses on constraining the second term of Variational Autoencoders loss function: The Kullback-Leibler component (Beta-VAE, FactorVAE Beta-TCVAE). Despite promising results, this comes with a trade-off between disentanglement and reconstruction. Besides, it is not clear why minimizing the KL divergence leads to disentanglement.

In this paper, we propose to achieve disentangled representations by sampling from a sparse distribution. To give a visual appealing reconstruction for humans, we replace the conventional pixel-wise quadratic by perceptual loss. We demonstrate the reconstruction quality and disentangled on synthetic datasets.

## 1. Introduction

The success of machine learning algorithms relies deeply on the representations of the data. There is a consensus that good representations are distributed, invariant, and disentangled.

Disentangled representations assume that different neurons present in the latent space are uncorrelated, i.e., each is trying to learn something different about the input data. [Higgins 17]; [Kim 18] and [Eastwood 18] define a disentangled representation as a representation where a change in a single generative factor leads to a change in a single factor in the learned representation.

Learning disentangled representation is said to be useful for downstream tasks that are trying to learn, since they should contain all the information present in the data in an interpretable and understandable structure [Gondal 19, Bengio 13, Chen 17, Whitney 16].

A particular domain in which disentanglement can be applied is in reinforcement learning. Because the agent learns from sparse reward, the VAE framework can be used as a feature extractor. We can train the agent on compressed data representations instead of input space. From these lower representations, the network can extract useful causal features.

A state-of-the-artwork on learning disentangled representations $\beta$-VAE [Higgins 17] is based on Variational Autoencoders framework [Kingma 13]. The model focuses on constraining the latent space by imposing a large weight $\beta > 1$ to the Kl divergence between the posterior and prior distribution. Despite promising results, it comes with a trade-off between reconstruction accuracy and disentanglement. The reconstructed image is poor compared to conventional VAE [Kingma 13].

In this work, we aim to achieve disentanglement. We work with the conventional Variational Autoencoder framework(VAE) with a modified encoder network. We place a penalty on the encoder distribution. This encourages the model to sample from sparse distribution. Sampling from sparse distribution is said to be beneficial, since, the model learns salient features of the data, besides, it also reduces the risk of overfitting.

Since $\beta$-VAE reconstruction relies on pixel-wise quadratic, we address a limitation that may be caused by pixel-wise by replacing with a perceptual loss.

Our experiments on the celebrity and dSprites dataset demonstrate that this approach achieves a disentanglement. Although, not superior to $\beta$-VAE. We replace conventional wise-pixel quadratic by perceptual loss. The reconstruction quality is appealing to human eyes on CelebA than $\beta$-VAE and VAE.

Contact: cristovao-paulino@aist.go.jp

## 2. Background

The VAE model [Kingma 13] is powerful framework for learning explanatory factors of variations in the data. Its loss function Eq.1 has two components: the reconstruction and Kullback-Leibler divergence term. Despite good performance on image classification [Salimans 15], image segmentation [Sohn 15], text generation [Bowman 15], and artistic applications [Chan 19], VAE loss function does not encourage any structure on the latent space.

$$L(\theta, \phi) = \mathbb{E}_{z\sim q_\theta(z|x)}[logp_\phi(x|z)] - D_{KL}(q_\theta(z|x)||p(z)) \quad (1)$$

A state-of-the-artwork on disentanglement [Higgins 17], achieves disentanglement by introducing a scalar hyperparameter to the loss function Eq.2.

$$L(\theta, \phi) = \mathbb{E}_{z\sim q_\theta(z|x)}[logp_\phi(x|z)] - \beta D_{KL}(q_\theta(z|x)||p(z)) \quad (2)$$

This hyperparameter $\beta$ weigh Kullback-Leibler divergence (KL) present in the prior distribution, besides, penalizes the KL term, which encourages similarity on the factorized prior distribution. However, it is not yet explicit how do disentangled representation work.

Recently, there have been many works targeting disentanglement [Burgess 18, Chen 18, Kim 18]. They all introduce a hyperparameter to minimizing the KL divergence term. A different approach is taking by [Chen 16]; the model is

based on Generative Adversarial Neural Networks. It maximizes the mutual information between a small subset of the latent variables and the observation.

We seek to demonstrate an alternative approach to achieve disentanglement without penalizing the KL divergence.

## 3. Proposed Model

### 3.1 Overview

We induce sparsity at the latent space using the l1 norm at the encoder network, and we tackle the drawback of pixel-wise quadratic with a perceptual similarity loss that measures the latent representations.

In the following section, we explain our approaches to enhance disentangled representation and reconstruction quality.

### 3.2 Perceptual loss

Mean squared error (MSE) is often used as a reconstruction metric; it is easily implemented and is proven to be efficient for computer vision tasks. Several kinds of research on disentanglement employ pixel-wise quadratic metric. However, often the generated images tend to be blurry when compared to natural images, also it might come with some errors. This is since the pixel-to-pixel loss does not capture the perceptual difference and spatial correlation between images.

We replace MSE by a perceptual loss [Johnson 16] to preserve the details of the images and to generate a visually appealing image for human eyes. The perceptual loss compares high-level differences, such as content and style discrepancies, between images. We use a pre-trained VGG16 model[Simonyan 14] trained on ImageNet to extract features.

## 4. Enforce Sparsity at the Latent Codes

The related work tackle disentanglement by focusing on the Kullback-Leibler term since it has a beneficial self-regularizing effect, forcing the model to focus on the most essential and disentangled features.

Our simple premise is that if the latent space distribution is well structured, the model is able to disentangled factors of variations. Since sampling from sparse distribution forces the model to focus on essential features of the data and also decreases the risk of overfitting. Besides, it induces less correlation used by latent variables.

At the encoder network, we induce sparsity by placing an L1 penalty at the output. The L1 penalty forces the distribution to have smaller weights and more numbers of zeros. This is equivalent to have a few numbers of neurons active in the network.

Several theoretical, computational, and experimental studies suggest that neurons encode sensory information using a small number of active neurons at any given point in time. This strategy, referred to as sparse coding, could confer several advantages.

Table 1: reconstruction-dSprites

| | **PSRN** | **SSIM** |
|---|---|---|
| VAE | 64.2544 | 0.9184 |
| Beta-VAE | 64.2543 | 0.9180 |
| Ours | 64.2544 | 0.9184 |

Table 2: reconstruction-celebA

| | **PSRN** | **SSIM** |
|---|---|---|
| VAE | 58.4036 | 0.3126 |
| Beta-VAE | 58.4037 | 0.3126 |
| Ours | 54.9790 | 0.1545 |

In summary, we assume that if the latent space is well structured, salient features of variations on the data might be discovered.

## 5. Experiments

We analyze the disentanglement properties of the model qualitatively and quantitatively, through well-known datasets such as dSprites, CelebA, and 3d chairs.

We include the regularization term l1 penalty to encourage sparsity at the prior distribution and deep feature loss describe beforehand. We observe that these sparse latent variables, in fact, enable interpretability of the latent codes and encourages the model to disentangle factors of variation in the data such as rotation, scale, position in the dSprites dataset, and hairstyle, face orientation, gender, smile and age in the CelebA dataset.

### 5.1 Reconstruction

We quantitatively (PSRN and SSIM) and qualitative evaluate the Reconstruction. As shown in Fig.1, the model reconstruct the input data of dSprites. In Fig.2, is shown the reconstruction for VAE, $\beta$-VAE, and our proposed model. Initially, we trained the model with MSE and perceptual loss individually. Finally, simultaneously we trained the model with perceptual and sparse encoder. The effects of perceptual loss are visible in CelebA in Fig.2 [d and e]. Our reconstructed images are less blurry than other models and have a visual appealing for human eyes.

Surprisingly the evaluation metric is similar to other models on dSprites as shown in Table.1. The expected scenario is illustrated in 2, where the score is lower than $\beta$-VAE and VAE.

### 5.2 Disentanglement

In all figures is shown, the latent code traversal. Each block corresponds to the traversal of a single latent variable while keeping others fixed. There is no agreement on how to evaluate disentanglement, we adopted the Mutual Information gap (MIG) [Chen 18] and Axis Alignment Metric (AMM)[Dubois 19]. The model is considered to achieve a perfect disentanglement when MIG or AMM is 1.

The dSprites dataset is currently the only available dataset to evaluate quantitatively disentanglement. We summarize the results on Table 3. $\beta$-VAE achieves a higher score on MIG, surprisingly VAE achieves a higher score on

(a) Beta-vae (b) ours

Figure 1: Reconstruction - dSprites

(a) vae (b) beta-vae

(c) ours-mse (d) ours-perceptual

(e) ours-sparse+perceptual

Figure 2: Reconstruction - CelebA

Table 3: Disentanglement-dSprites

| | **MIG** | **AAM** |
|---|---|---|
| VAE | 1.6e-8 | 0.20 |
| Beta-VAE | 0.4 | 2.75e-8 |
| Ours | 0.2 | 2.36e-8 |

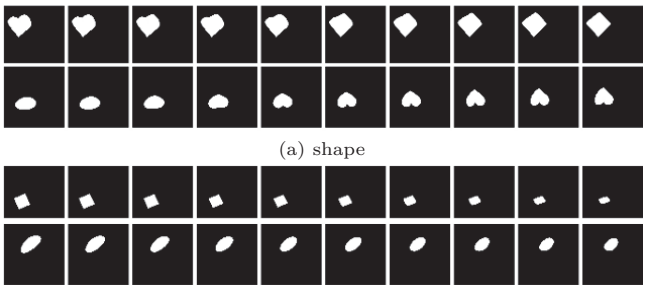
AMM. This requires further experiment, since the AMM is considerably lower for $\beta$-VAE and our model.

Inspecting the latent traversal in Fig.3, our model discovers the following factors: shape, scale, and scale. However, it does not discover the factor which generates the Y-position. The model fails to traverse from the X to the Y position. CelebA is not ideal to evaluate disentanglement since many factors of variation can be inferred.

However, researchers use the data to understand and interpret disentanglement. As shown in Fig.4, our model discovers in unsupervised manner factors that encode smile, hairstyle, a transition from male to female, rotation, and transition from an elderly male to younger. Our model achieves a certain level of disentanglement, it lacks more variation on the generated samples.

## 6. Conclusion

There are many possible ways to express the preference for learning disentangled representations as an objective. The related work proposes to penalize the KL divergence term. In this work, we propose to learn controllable factors



(a) shape

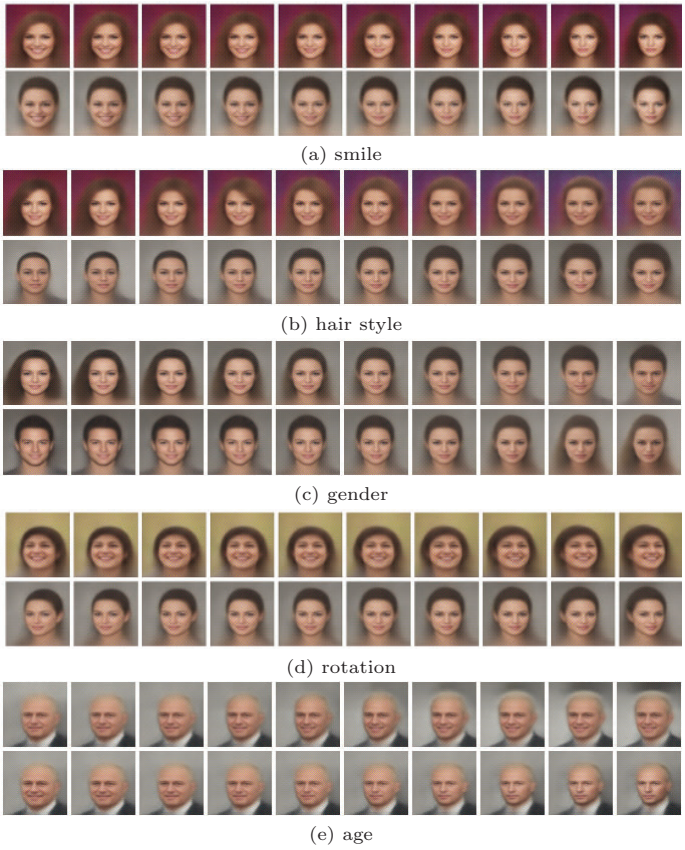
(b) scale

Figure 3: dSprites dataset - learned disentangled representation

of variation by enforcing sparsity in the latent space. We replace the conventional encoder by a sparse encoder network. Besides, we use perceptual loss at the reconstruction term.

We demonstrate that this approach achieves a certain level of disentanglement. The model discovers the factor of variations present in the data.

Further research is needed to achieve the same level as $\beta$-VAE. This may include the importance of weighted sampling and evaluate the disentanglement with methods purposed by [Higgins 17],[Kim 18].



(a) smile

(b) hair style

(c) gender

(d) rotation

(e) age

Figure 4: CelebA dataset - learned disentangled representation

Figure 5: 3d chairs dataset - learned disentangled representation

## Acknowledgement

This paper is based on results obtained from a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO). This work was supported by JSPS KAKENHI Grant Number 19K11994.

## References

[Bengio 13] Bengio, Y., Courville, A., and Vincent, P.: Representation learning: A review and new perspectives, *IEEE transactions on pattern analysis and machine intelligence*, Vol. 35, No. 8, pp. 1798–1828 (2013)

[Bowman 15] Bowman, S. R., et al.: Generating sentences from a continuous space, *arXiv preprint arXiv:1511.06349* (2015)

[Burgess 18] Burgess, C. P., et al.: Understanding disentangling in VAE, *arXiv preprint arXiv:1804.03599* (2018)

[Chan 19] Chan, C., et al.: Everybody dance now, in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5933–5942 (2019)

[Chen 16] Chen, X., et al.: Infogan: Interpretable representation learning by information maximizing generative adversarial nets, in *Advances in neural information processing systems*, pp. 2172–2180 (2016)

[Chen 17] Chen, X., et al.: Long-term video interpolation with bidirectional predictive network, in *Visual Communications and Image Processing (VCIP), 2017 IEEE*, pp. 1–4IEEE (2017)

[Chen 18] Chen, T. Q., et al.: Isolating Sources of Disentanglement in Variational Autoencoders, *arXiv preprint arXiv:1802.04942* (2018)

[Dubois 19] Dubois, Y., Kastanos, A., Lines, D., and Melman, B.: Understanding Disentangling in VAE (2019)

[Eastwood 18] Eastwood, C. and Williams, C. K.: A framework for the quantitative evaluation of disentangled representations (2018)

[Gondal 19] Gondal, M. W., et al.: On the Transfer of Inductive Bias from Simulation to the Real World: a New Disentanglement Dataset, *arXiv preprint arXiv:1906.03292* (2019)

[Higgins 17] Higgins, I., et al.: beta-vae: Learning basic visual concepts with a constrained variational framework, in *International Conference on Learning Representations* (2017)

[Johnson 16] Johnson, J., et al.: Perceptual losses for real-time style transfer and super-resolution, in *European conference on computer vision*, pp. 694–711Springer (2016)

[Kim 18] Kim, H. and Mnih, A.: Disentangling by factorising, *arXiv preprint arXiv:1802.05983* (2018)

[Kingma 13] Kingma, D. P. and Welling, M.: Auto-encoding variational bayes, *arXiv preprint arXiv:1312.6114* (2013)

[Salimans 15] Salimans, T., Kingma, D., and Welling, M.: Markov chain monte carlo and variational inference: Bridging the gap, in *International Conference on Machine Learning*, pp. 1218–1226 (2015)

[Simonyan 14] Simonyan, K. and Zisserman, A.: Very deep convolutional networks for large-scale image recognition, *arXiv preprint arXiv:1409.1556* (2014)

[Sohn 15] Sohn, K., et al.: Learning structured output representation using deep conditional generative models, in *Advances in neural information processing systems*, pp. 3483–3491 (2015)

[Whitney 16] Whitney, W. F., et al.: Understanding visual concepts with continuation learning, *arXiv preprint arXiv:1602.06822* (2016)