Towards Disentangled Representations via Variational Sparse Coding

Robert Aduviri and Alfredo De La Fuente

Pontifical Catholic University of Peru, Skolkovo Institute of Science and Technology

Abstract

We present a framework for learning disentangled representations with variational autoencoders in an unsupervised manner, which explicitly imposes sparsity and interpretability of the latent encodings. Leveraging ideas from Sparse Coding models, we consider the Spike and Slab prior distribution for the latent variables, and a modification of the ELBO, inspired by the β-VAE model to enforce decomposability over the latent representation. We run our proposed model in a variety of quantitative and qualitative experiments for MNIST, Fashion-MNIST, CelebA and dSprites datasets, showing that the framework disentangles the latent space in continuous sparse interpretable factors and is competitive with current disentangling models.

Introduction

- Learning **interpretable factorized representations** of data without supervision is regarded as an open challenge with important consequences for machine learning research.
- Disentanglement constitutes the *complex* task of learning representations that separate the **underlying structure** of the world into **disjoint parts of its representation**.

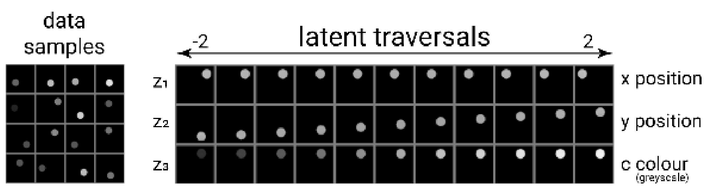

Figure 1: Scheme from the paper "Towards a Definition of Disentangled Representations" by Higgins et al. (2018)

β - VAE and the dSprites dataset

- β **- VAE:** Proposed by Higgins et al. [1] as a constrained version of VAE to discover disentangled latent factors.

$$\mathcal{L}_{\text{Beta}}(\theta, \phi) = \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] - \beta\, \mathbb{KL}\left(q_\phi(z|x) \| p(z)\right)$$

- **dSprites dataset**: Created by Matthey et al. [2] as a way to assess the disentanglement properties of unsupervised learning methods. These 2D shapes were procedurally generated from 6 ground truth independent latent factors: color, shape, scale, rotation, x and y positions.

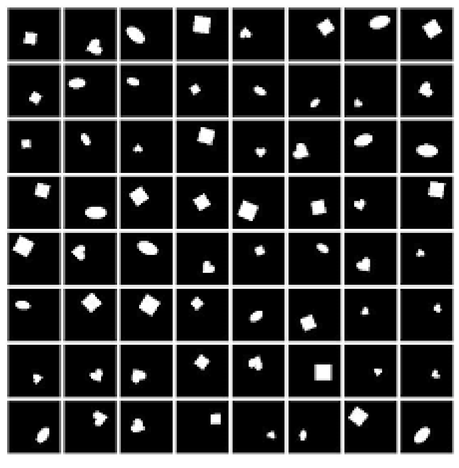

Figure 2: The dSprites dataset

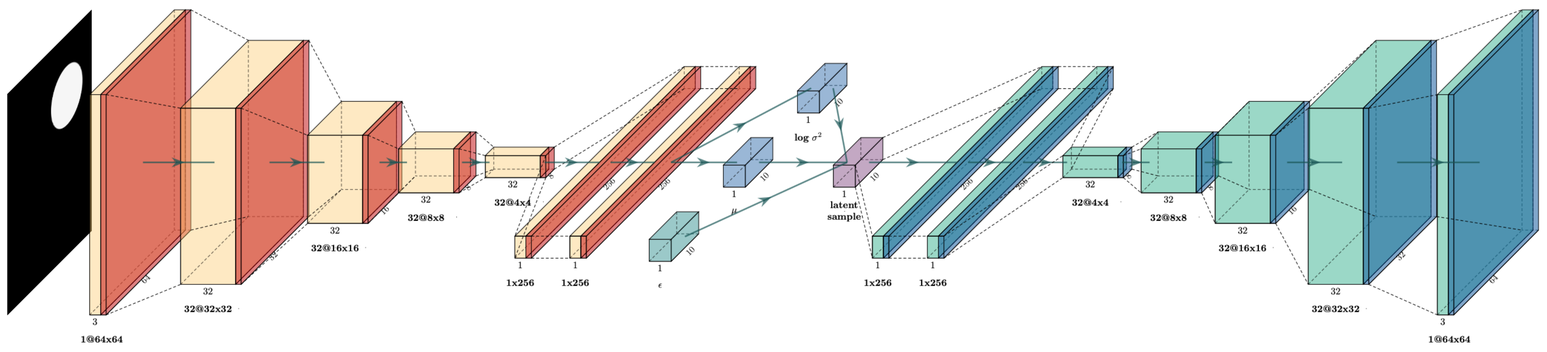

Figure 3: Architecture of the Convolutional Variational Sparse Coding model

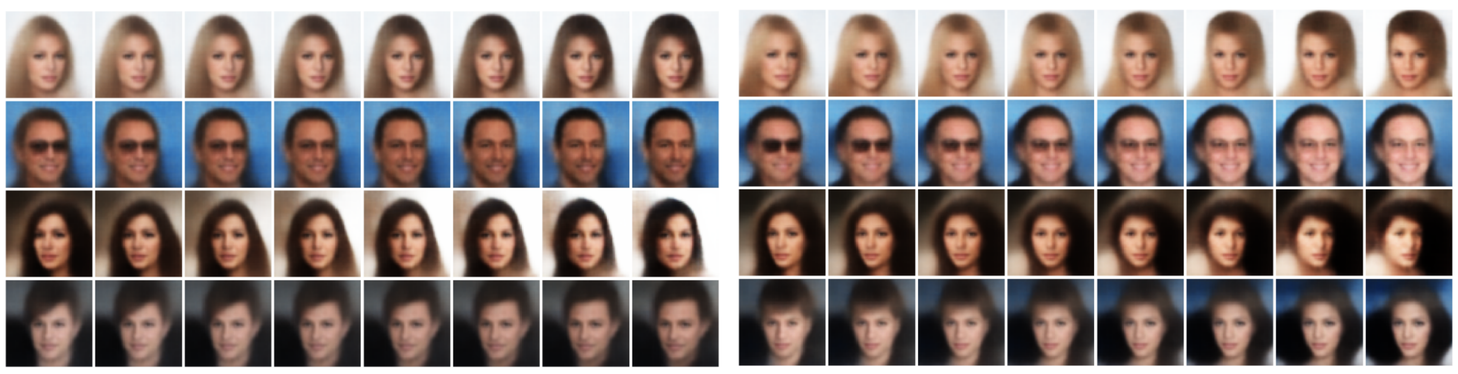

Figure 4: Latent traversals using the Convolutional VSC (left) and Convolutional VAE (right) models with the CelebA dataset

Variational Sparse Coding

- The Variational Sparse Coding (VSC) model [3], consists in a variational autoencoder model (VAE) with a **Spike and Slab** prior, which induces **sparsity in the learned latent space** (Figure 5).
- We focus our analysis on assessing the disentanglement properties of the VSC model beyond qualitative analysis through well-known datasets such as CelebA, incorporating quantitative disentanglement metrics.
- We also include the regularization term proposed by β-VAE and improvements as suggested in our previous reproducibility work [4].

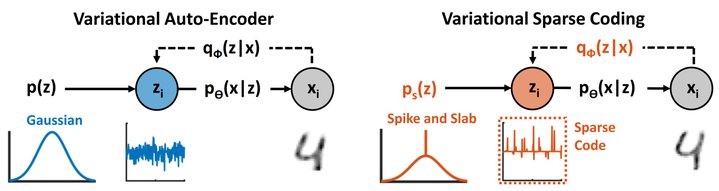

Figure 5: The VSC model induces sparsity in the latent space via a Spike and Slab distribution

Results

- The Spike and Slab distribution effectively induces sparsity in the latent codes with no detriment to the reconstruction quality (Figure 6).
- We observe that these sparse latent codes indeed represent interpretable and nearly independent variations in the generated images, such as shape, rotation, size and position in the dSprites dataset, and hair color, glasses, face orientation and hair bangs in the CelebA dataset (Figures 4 and 8).
- The traversal of the latent space is performed varying the latent codes with a high absolute value for a given image, one at a time.

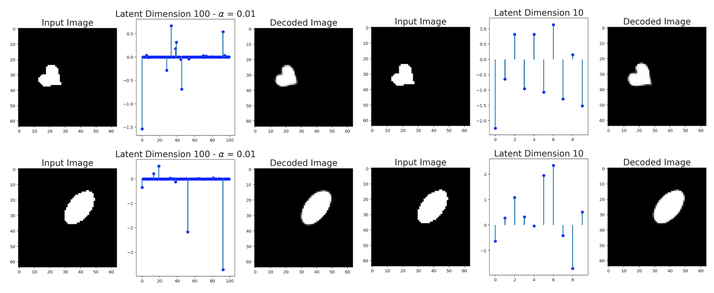

Figure 6: Reconstruction and latent codes of Convolutional VSC (left) ($\alpha = 0.01$, $\beta = 2$) and Convolutional VAE (right) ($\beta = 2$) models with the dSprites dataset

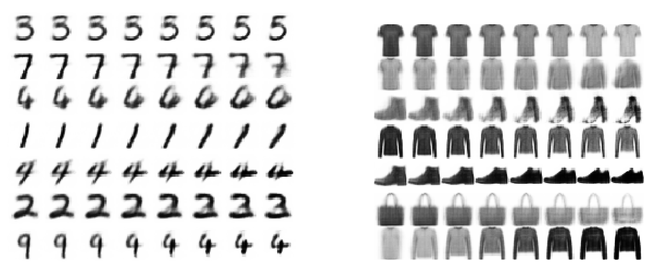

Figure 7: Latent traversals on MNIST (left) and Fashion-MNIST (right).

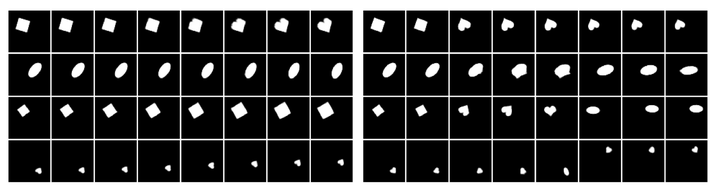

Figure 8: Latent traversals using the Convolutional VSC (left) and Convolutional VAE (right) models with the dSprites dataset

Next steps

- We are currently evaluating the results with recent disentanglement metrics to provide a quantitative evaluation of disentanglement to be compared to related work, such as:
  - **Metrics:** BetaVAE score, FactorVAE score, Mutual Information Gap, SAP score, DCI, MCE, IRS
  - **Models:** FactorVAE, BetaTCVAE, DIP-VAE, InfoGAN
  - **Datasets:** Color/Noisy/Scream-dSprites, SmallNORB, Cars3D, Shapes3D
- The recent disentanglement_lib created by Locatello et al. [5] presents as a valuable framework for conducting quantitative disentanglement experiments.

References

[1] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. β-vae: Learning basic visual concepts with a constrained variational framework.
In *ICLR 2017*, volume 3, 2017.

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

[3] Francesco Tonolini, Bjorn Sand Jensen, and Roderick Murray-Smith. Variational Sparse Coding, 2019.

[4] Alfredo de la Fuente Briceño and Robert Aduviri. Alfo5123/Variational-Sparse-Coding: First Release of Submission for ICLR Reproducibility Challenge 2019, May 2019.

[5] 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.