Disentanglement

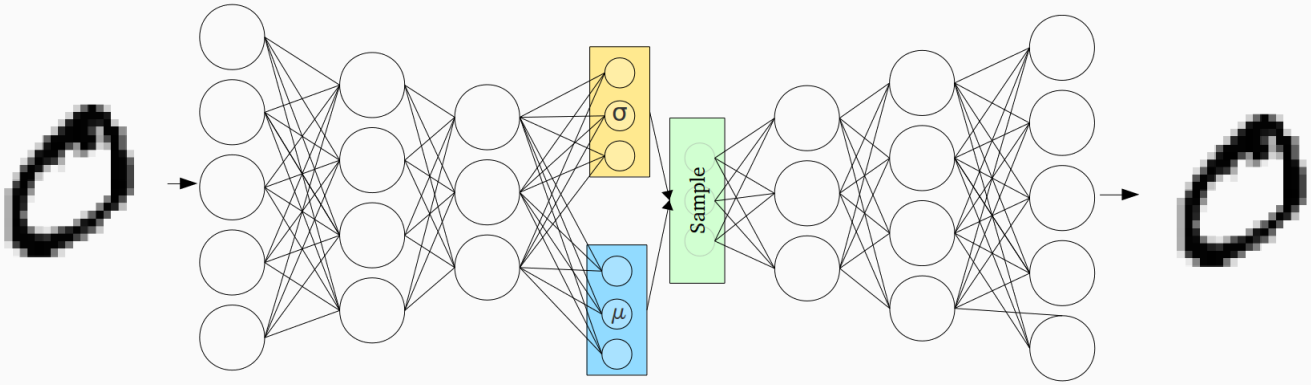
Matthew Painter

May 8, 2019

University of Southampton

Generative Models

VAE



GAN

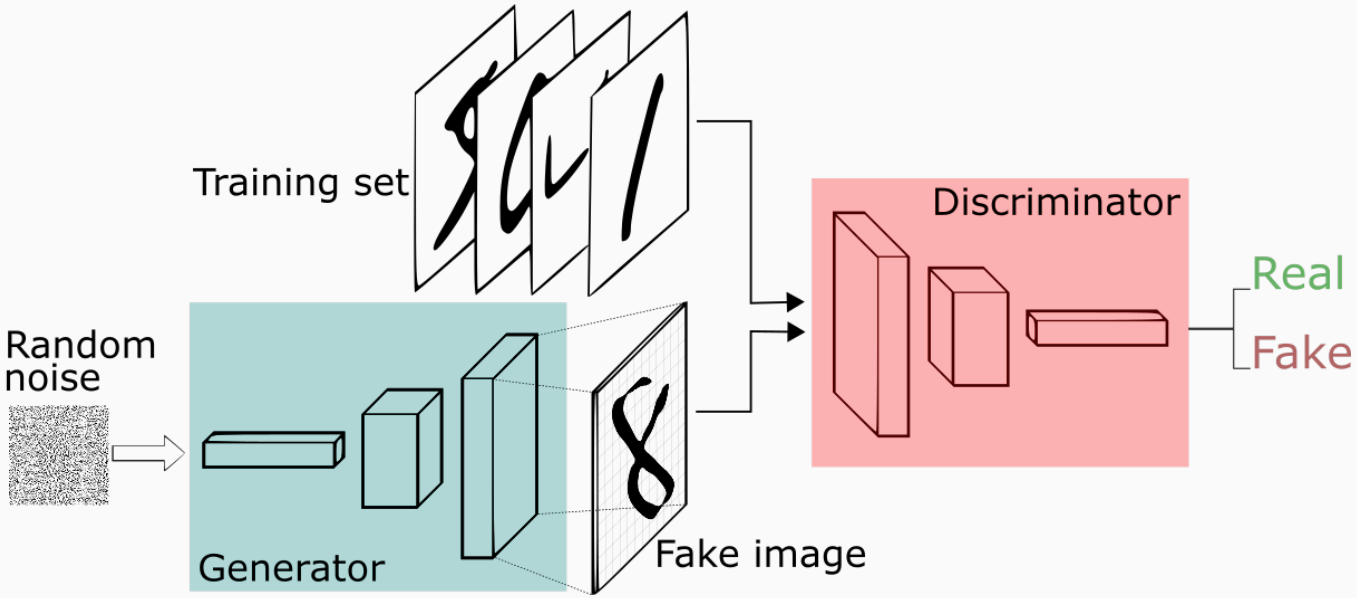


Image (top): skymind.ai/wiki/generative-adversarial-network-gan

Image (bottom): towardsdatascience.com/intuitively-understanding-variational-autoencoders-1bfe67eb5daf

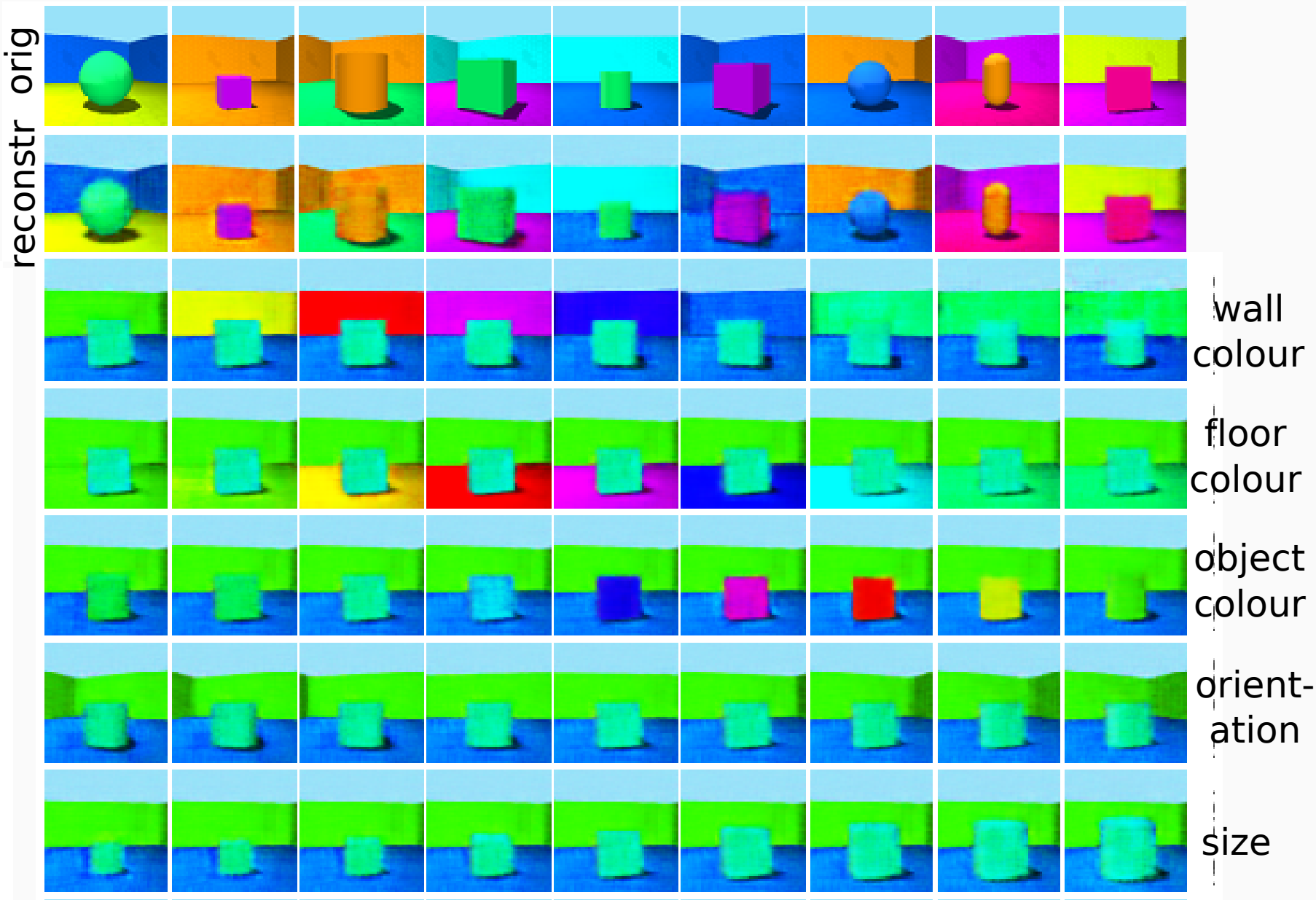
Disentanglement - Definitions

Beyond being distributed and invariant, we would like our representations to disentangle the factors of variation. Different explanatory factors of the data tend to change independently of each other in the input distribution, and only a few at a time tend to change when one considers a sequence of consecutive real-world inputs. **Bengio et al. [2013]**

A disentangled representation is one for which changes in the encoded data are sparse over real world transformations; that is, changes in only a few latents at a time should be able to represent sequences which are likely to happen in the real world. **Kulkarni et al. [2015]**

A disentangled representation can be defined as one where single latent units are sensitive to changes in single generative factors, while being relatively invariant to changes in other factors. **Higgins et al. [2017]**

Disentanglement - Visually



Disentanglement - Why?

Why do we want disentangled representations?

- Aesthetics
- Efficient
- Analogous to human visualisation

Deep Convolutional Inverse Graphics

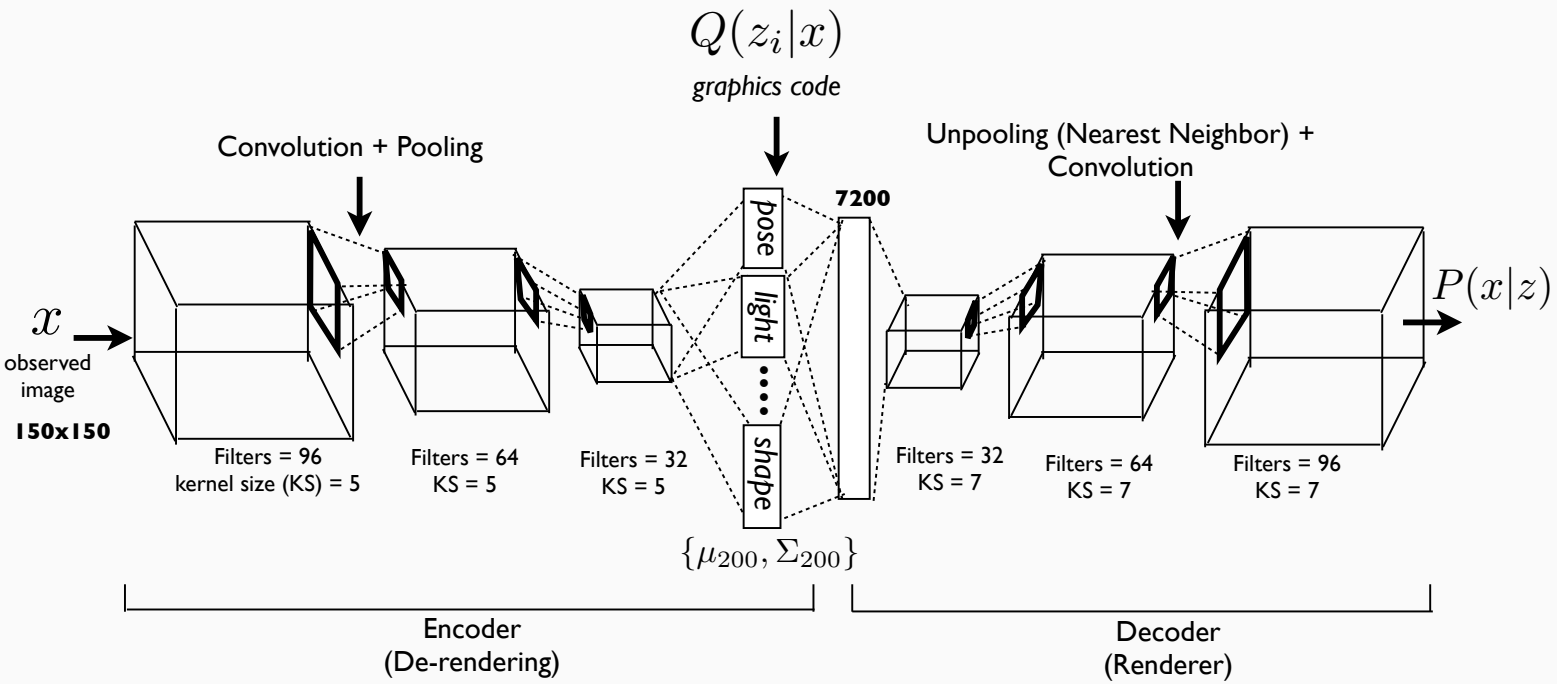


Figure 1: DC-IGN Model [Kulkarni et al., 2015]

Deep Convolutional Inverse Graphics - Results

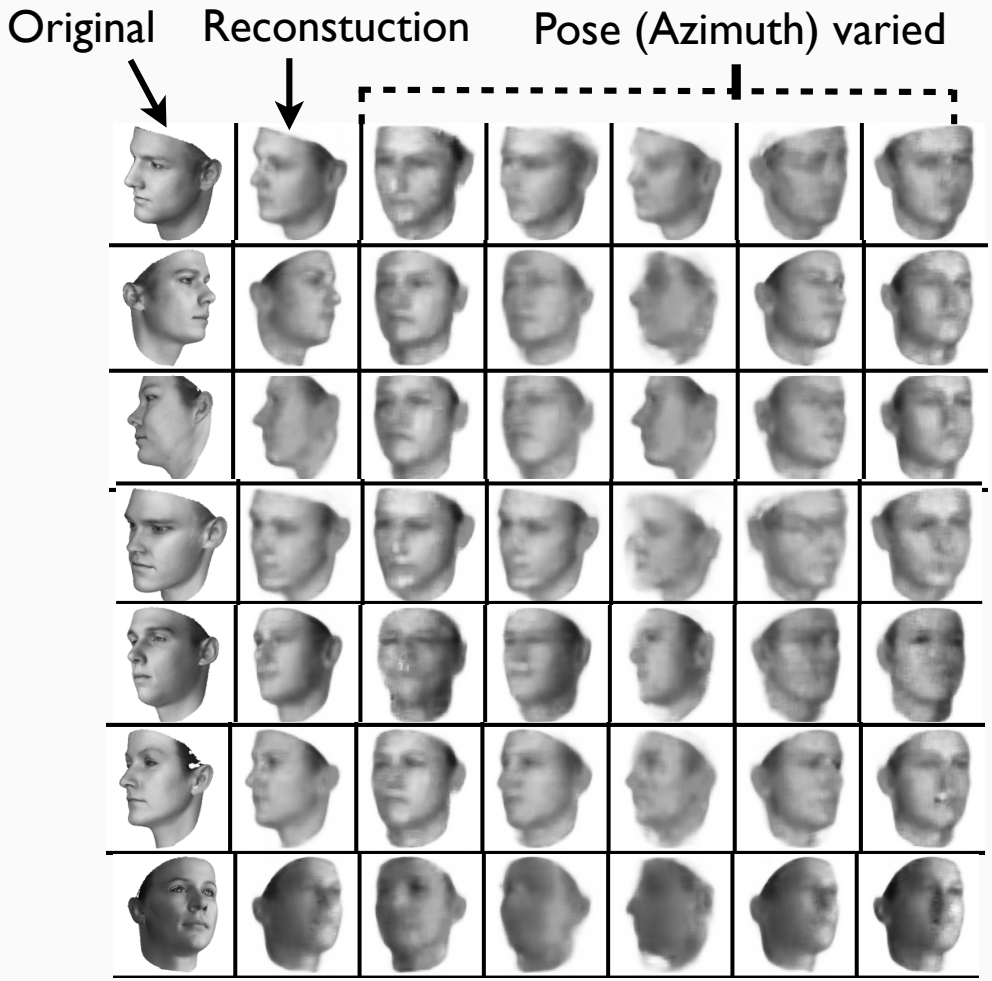


Figure 2: DC-IGN Vary Azimuth [Kulkarni et al., 2015]

β-VAE

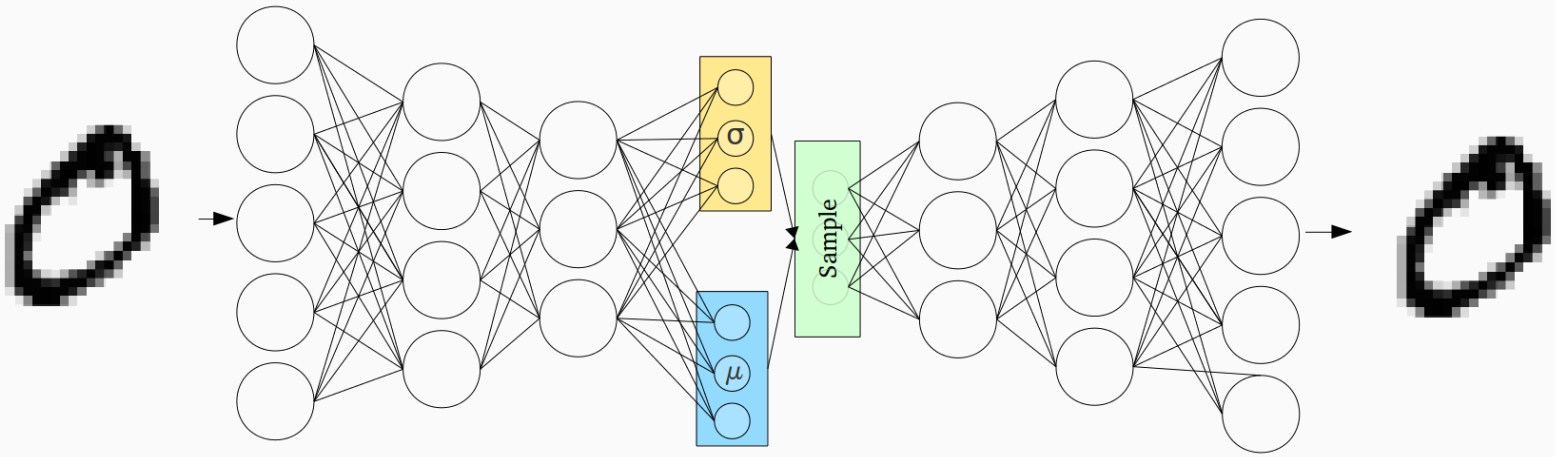


Figure 3: β-VAE Model [Higgins et al., 2017]

$$\mathcal{L} = \underbrace{\mathbb{E}_{q_\phi}[\log p_\theta(\mathbf{x}|\mathbf{z})]}_{\text{Reconstruct well}} - \underbrace{\beta D_{KL}\left(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})\right)}_{\text{Latents} \sim \text{normal gaussian}}, \quad \beta \in \mathbb{R}_+$$

β-VAE - Results

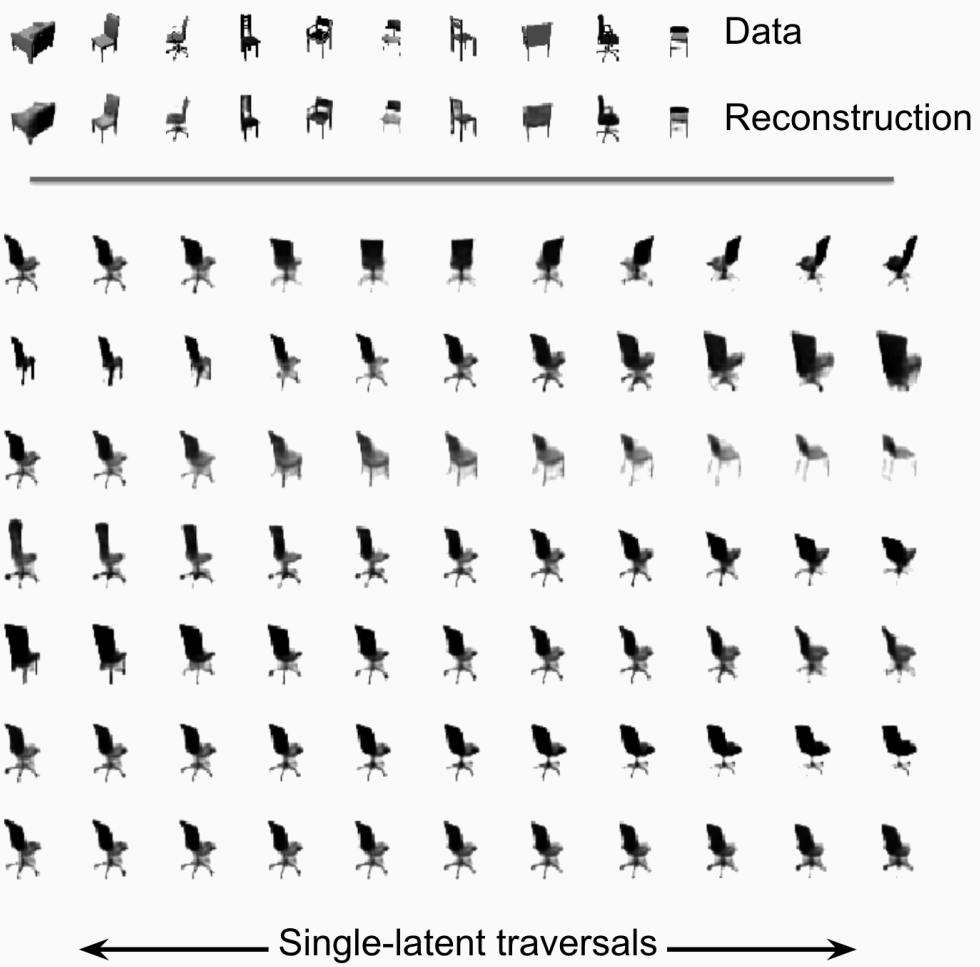


Figure 4: β-VAE Results [Higgins et al., 2017]

Metrics

There are a number of proposed metrics:

- Higgins et al. [2017] - Fix one factor and try to classify which factor was fixed.
- Kim and Mnih [2018] - Fix one factor and find the minimally varying latent dimension. Classify from that index to fixed factor.
- Chen et al. [2018] - Measure mutual information between latent dimension and generative factor. Measure difference in this for top two latent dimensions for each factor.

Metrics - Comparison

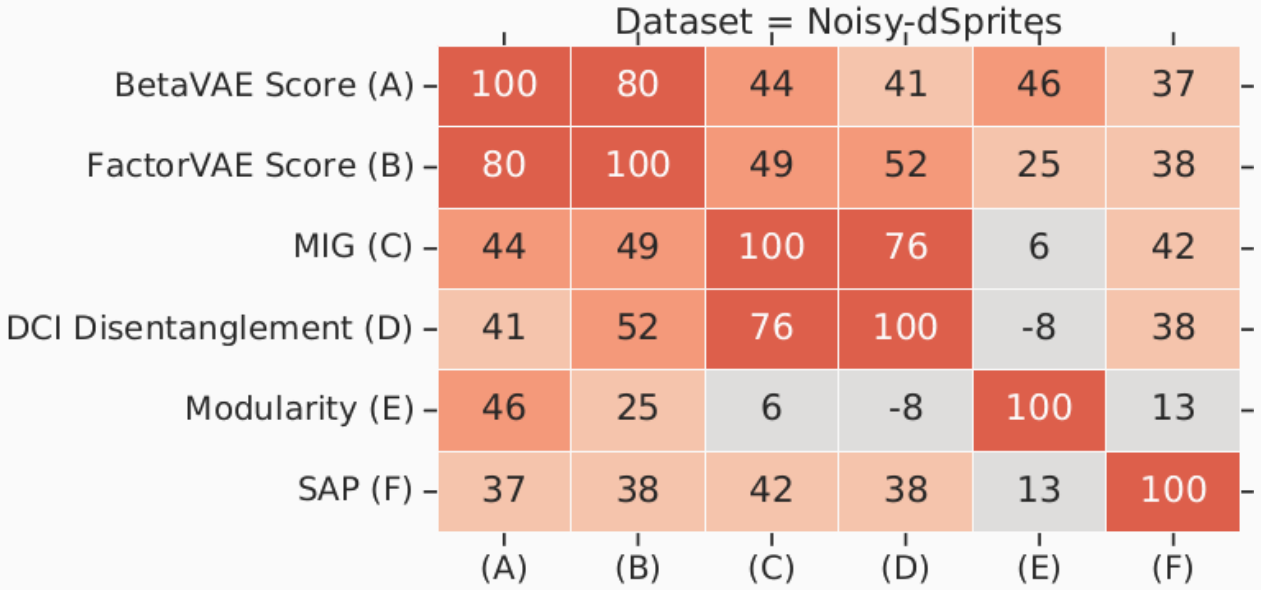


Figure 5: Spearman rank correlation between metrics by Locatello et al. [2018]

Papers

- beta-vae: learning visual concepts with a constrained variational framework
- Disentangling by Factorising
- Understanding disentangling in beta-vae
- Isolating Sources of Disentanglement in Variational Autoencoders
- Towards a Definition of Disentangled Representations
- InfoVAE: Information Maximizing Variational Autoencoders

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.

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.

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. In *International Conference on Learning Representations*, volume 3, 2017.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. *arXiv preprint arXiv:1802.05983*, 2018.

Tejas D Kulkarni, William F Whitney, Pushmeet Kohli, and Josh Tenenbaum. Deep convolutional inverse graphics network. In *Advances in neural information processing systems*, pages 2539–2547, 2015.

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.