Disentangling Disentanglement in Variational Autoencoders

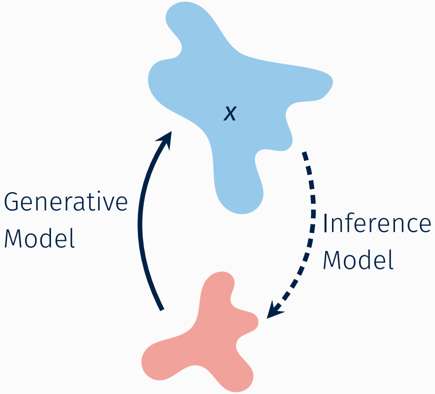
ICML 2019

Emile Mathieu*, Tom Rainforth*, N. Siddharth*, Yee Whye Teh

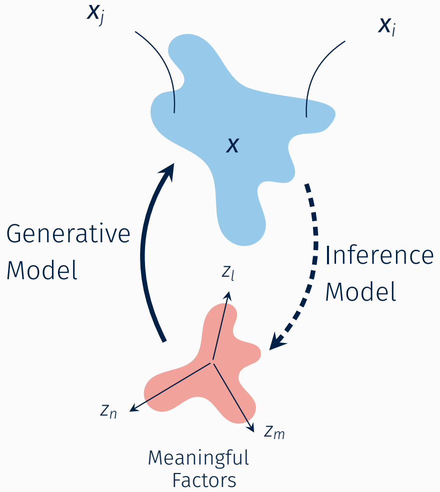
June 12, 2019

Departments of Statistics and Engineering Science, University of Oxford

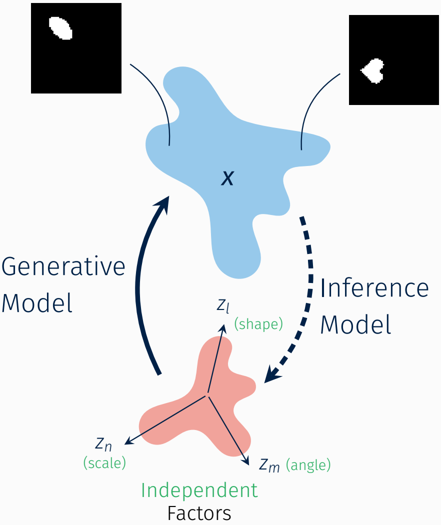
Variational Autoencoders



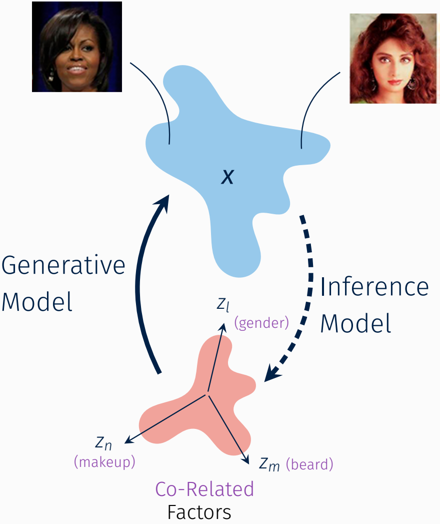
Disentanglement



Disentanglement = Independence



Decomposition $\in$ {Independence, Clustering, Sparsity, ...}



Decomposition: A Generalization of Disentanglement

Characterise decomposition as the fulfilment of two factors:

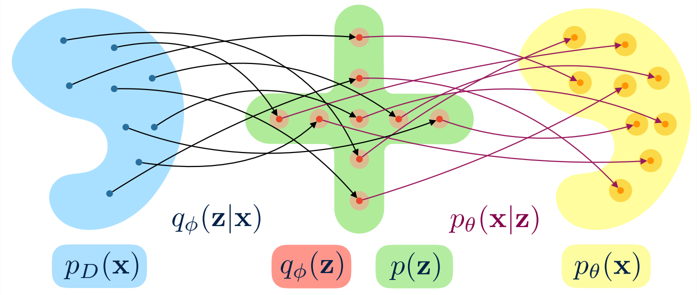
(a) level of overlap between encodings in the latent space,

(b) matching between the marginal posterior $q_\phi(\boldsymbol{z})$ and structured prior $p(\boldsymbol{z})$ to constrain with the required decomposition.

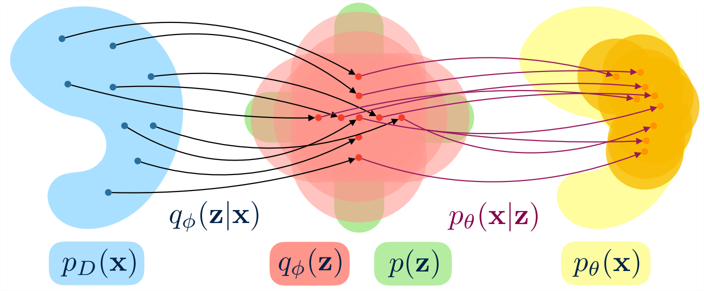
Desired Structure

$p(\mathbf{z})$

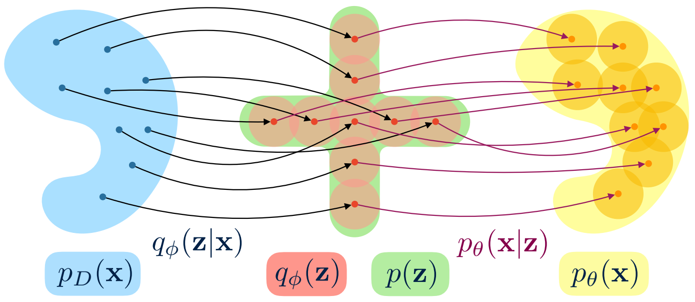
Insufficient Overlap



Too Much Overlap



Appropriate Overlap



Overlap — Deconstructing the β-VAE

$$\begin{aligned}\mathcal{L}_\beta(x) &= \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \beta \cdot \mathrm{KL}(q_\phi(z|x)||p(z)) \\ &= \underbrace{\mathcal{L}(x)\,(\pi_{\theta,\beta}, q_\phi)}_{\text{ELBO with } \beta\text{-annealed prior}} + \underbrace{(\beta - 1) \cdot H_{q_\phi}}_{\text{maxent}} + \underbrace{\log F_\beta}_{\text{constant}}\end{aligned}$$

Implications

β-VAE disentangles largely by controlling the level of overlap

It places no direct pressure on the latents to be independent!

Decomposition: Objective

$$\mathcal{L}_{\alpha,\beta}(x) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x \mid z)]$$ Reconstruct observations

$$- \beta \cdot \mathrm{KL}(q_\phi(z \mid x) \parallel p(z))$$ Control level of overlap

$$- \alpha \cdot \mathbb{D}(q_\phi(z), p(z))$$ Impose desired structure

Independence: $p(z) = \mathcal{N}(0, \boldsymbol{\sigma}^{\star})$

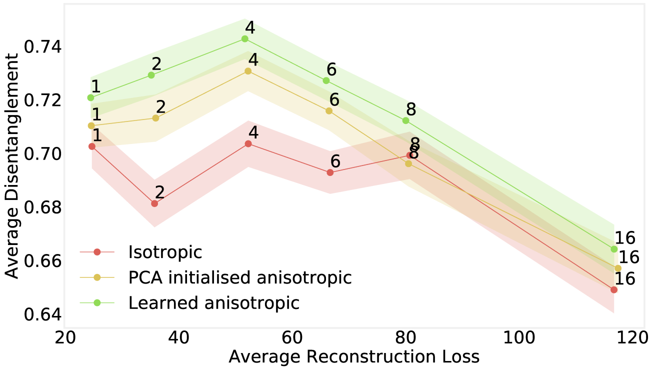


Figure 1: β-VAE trained on *2D Shapes*[1] computing disentanglement[2].

[1] Matthey et al., *dSprites: Disentanglement testing Sprites dataset*, p. 1.

[2] Kim and Mnih, “Disentangling by Factorising”, p. 2.

Decomposition: Generalising Disentanglement

Clustering: $p(\boldsymbol{z}) = \sum_k \rho_k \cdot \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k)$

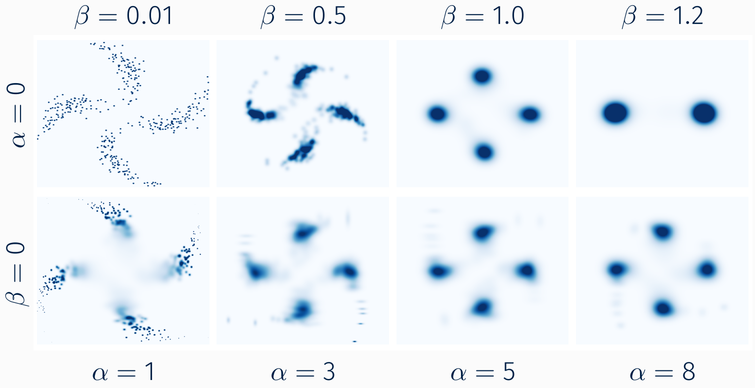


Figure 2: Density of aggregate posterior $q_\phi(\boldsymbol{z})$ with different α, β for the pinwheel dataset.[3]

[3] `http://hips.seas.harvard.edu/content/synthetic-pinwheel-data-matlab`.

Decomposition: Generalising Disentanglement

Sparsity: $p(\boldsymbol{z}) = \prod_d \; (1-\gamma)\cdot\mathcal{N}(z_d; 0, 1) + \gamma\cdot\mathcal{N}(z_d; 0, \sigma_0^2)$

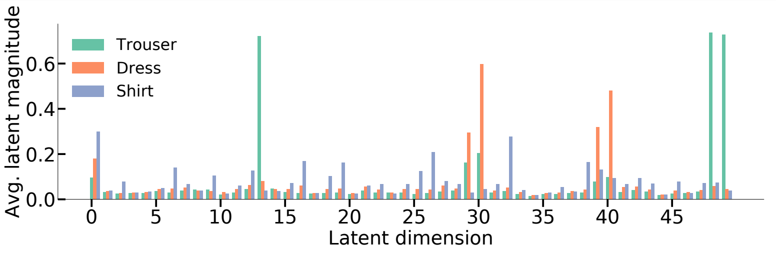
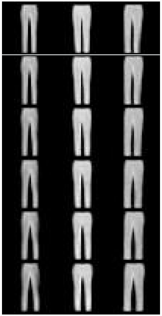


Figure 3: Sparsity of learnt representations for the *Fashion-MNIST*[4] dataset.

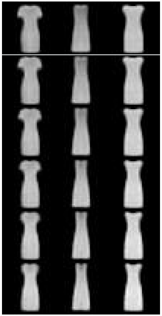
[4] Xiao, Rasul, and Vollgraf, *Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms.*

Decomposition: Generalising Disentanglement

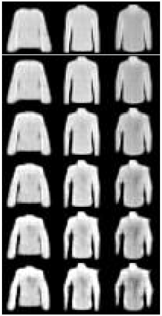
Sparsity: $p(\mathbf{z}) = \prod_d \; (1-\gamma) \cdot \mathcal{N}(\mathbf{z}_d; 0, 1) + \gamma \cdot \mathcal{N}(\mathbf{z}_d; 0, \sigma_0^2)$



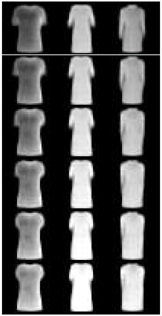
(a) $d = 49$ leg separation



(b) $d = 30$ dress width



(c) $d = 19$ shirt fit



(d) $d = 40$ sleeve style

Figure 3: Latent space traversals for "active" dimensions[4].

[4] Xiao, Rasul, and Vollgraf, *Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms*.

Sparsity: $p(z) = \prod_d \; (1-\gamma)\cdot\mathcal{N}(z_d; 0, 1) + \gamma\cdot\mathcal{N}(z_d; 0, \sigma_0^2)$

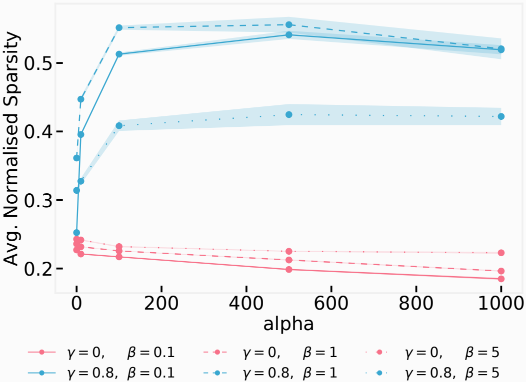


Figure 3: Sparsity vs regularisation strength α (higher better)[4].

[4] Xiao, Rasul, and Vollgraf, *Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms.*

Recap

We propose and develop:

- Decomposition: a generalisation of disentanglement involving:
  (a) overlap of latent encodings
  (b) match between $q_\phi(\boldsymbol{z})$ and $p(\boldsymbol{z})$
- A theoretical analysis of the β-VAE objective showing it primarily only contributes to overlap.
- An objective that incorporates both factors (a) and (b).
- Experiments that showcase efficacy at different decompositions:
  • independence • clustering • sparsity

Emile Mathieu

Tom Rainforth

N. Siddharth

Yee Whye Teh

Code

Paper

`iffsid/disentangling-disentanglement`

`arXiv:1812.02833`

Come talk to us at our poster: #5