Disentangling Disentanglement in Variational Autoencoders

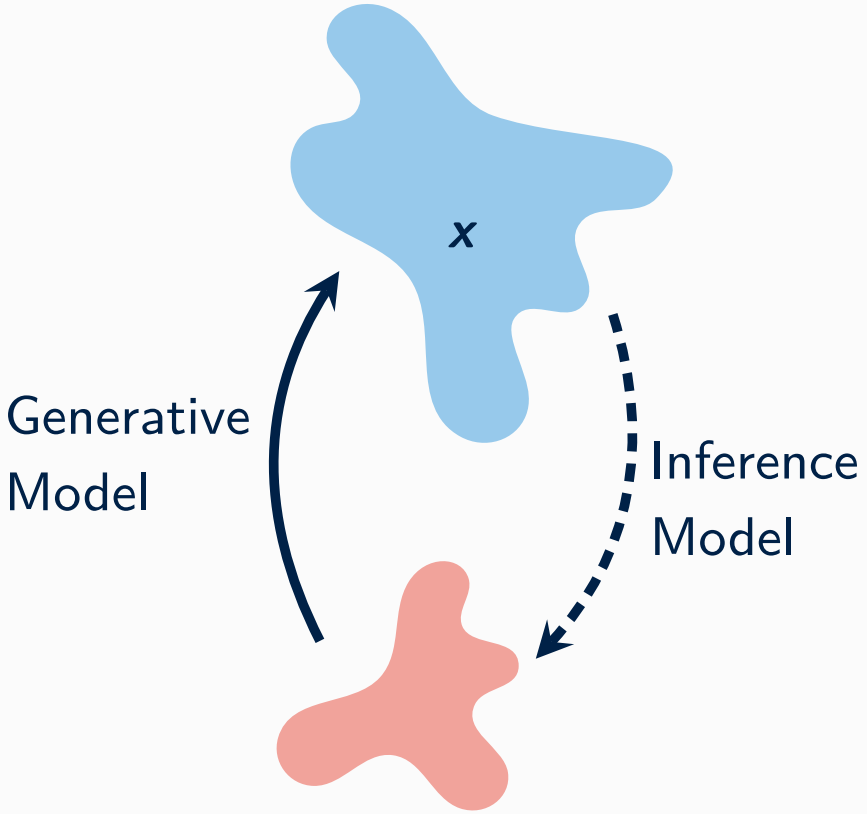
GenU 2019 (Originally Presented at ICML 2019)

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

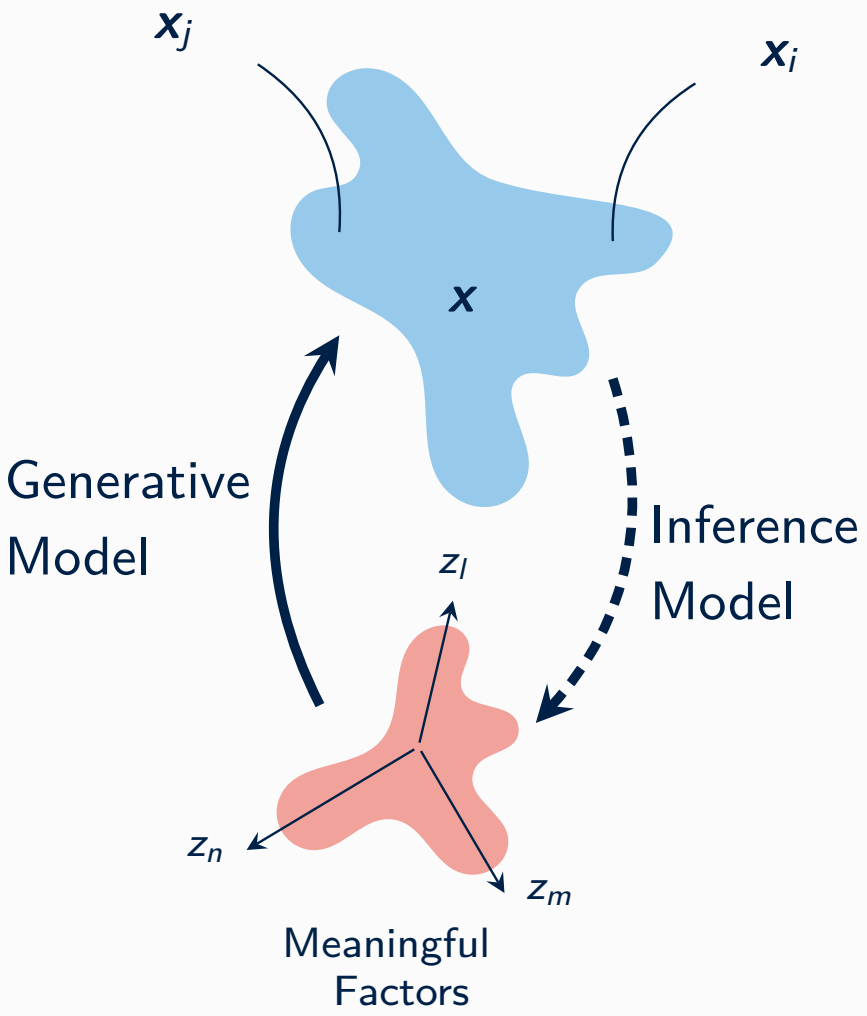
October 9th, 2019

Departments of Statistics and Engineering Science, University of Oxford

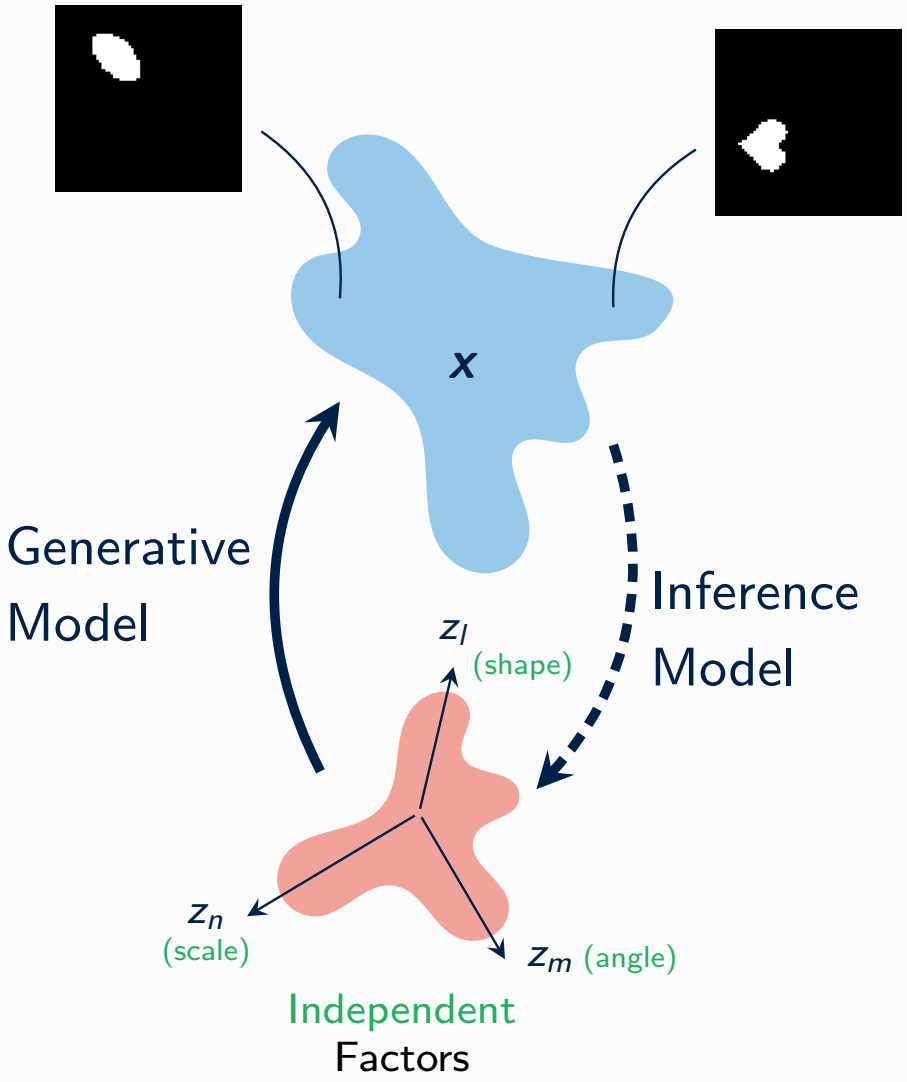
Variational Autoencoders



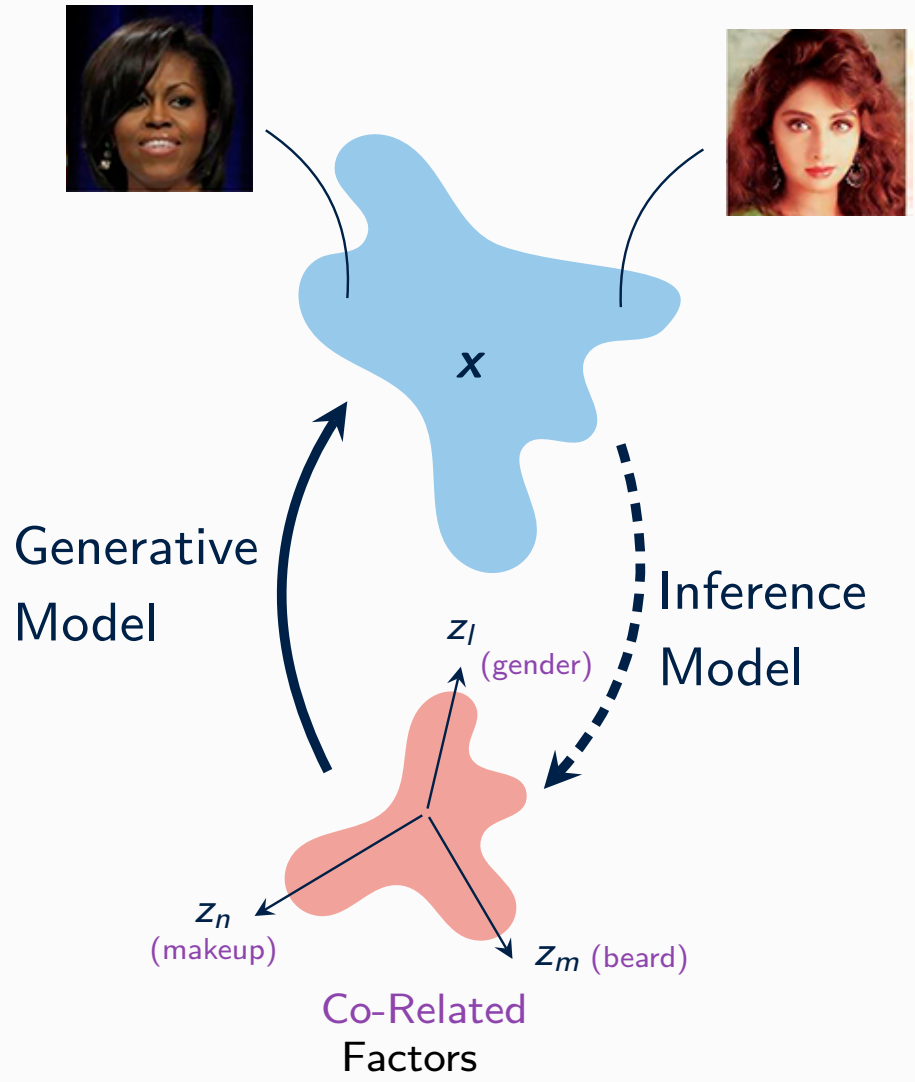
Interpretable Representations



Disentanglement = Independence



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



Talk Outline

- Background on VAEs and disentanglement

Talk Outline

- Background on VAEs and disentanglement
- Decomposition: a generalisation of disentanglement allowing a richer class of properties to be imposed

Talk Outline

- Background on VAEs and disentanglement
- Decomposition: a generalisation of disentanglement allowing a richer class of properties to be imposed
- Theoretical analysis of β-VAE

Talk Outline

- Background on VAEs and disentanglement
- Decomposition: a generalisation of disentanglement allowing a richer class of properties to be imposed
- Theoretical analysis of β-VAE
- An objective for encouraging decomposition

Talk Outline

- Background on VAEs and disentanglement
- Decomposition: a generalisation of disentanglement allowing a richer class of properties to be imposed
- Theoretical analysis of β-VAE
- An objective for encouraging decomposition
- Experiments that showcase examples decompositions:
  • independence • clustering • sparsity

Talk Outline

- Background on VAEs and disentanglement
- Decomposition: a generalisation of disentanglement allowing a richer class of properties to be imposed
- Theoretical analysis of β-VAE
- An objective for encouraging decomposition
- Experiments that showcase examples decompositions:
  • independence • clustering • sparsity
- Points of discussion

Variational Autoencoders (VAEs)

- Have observations $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ from data generating process $p_{\mathcal{D}}(\boldsymbol{x})$

Variational Autoencoders (VAEs)

- Have observations $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ from data generating process $p_{\mathcal{D}}(\boldsymbol{x})$
- Want to learn a corresponding latent variable model $p(\boldsymbol{z})p_\theta(\boldsymbol{x}|\boldsymbol{z})$

$$\arg\min_\theta \mathrm{KL}(p_{\mathcal{D}}(\boldsymbol{x}) \parallel p_\theta(\boldsymbol{x})) = \arg\max_\theta \mathbb{E}_{p_{\mathcal{D}}(\boldsymbol{x})}[\log p_\theta(\boldsymbol{x})]$$

Variational Autoencoders (VAEs)

- Have observations $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ from data generating process $p_{\mathcal{D}}(\boldsymbol{x})$
- Want to learn a corresponding latent variable model $p(\boldsymbol{z})p_\theta(\boldsymbol{x}|\boldsymbol{z})$

$$\arg\min_\theta \mathrm{KL}(p_{\mathcal{D}}(\boldsymbol{x}) \parallel p_\theta(\boldsymbol{x})) = \arg\max_\theta \mathbb{E}_{p_{\mathcal{D}}(\boldsymbol{x})}[\log p_\theta(\boldsymbol{x})]$$

- As this intractable, instead introduce inference model $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ and replace $\log p_\theta(\boldsymbol{x})$ with the lower bound

$$\begin{aligned}\mathcal{L}(\boldsymbol{x};\theta,\phi) &\triangleq \log p_\theta(\boldsymbol{x}) - \mathrm{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \parallel p_\theta(\boldsymbol{z}|\boldsymbol{x})) \\ &= \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - \mathrm{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \parallel p(\boldsymbol{z})).\end{aligned} \tag{1}$$

Variational Autoencoders (VAEs)

- Have observations $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ from data generating process $p_{\mathcal{D}}(\boldsymbol{x})$
- Want to learn a corresponding latent variable model $p(\boldsymbol{z})p_\theta(\boldsymbol{x}|\boldsymbol{z})$

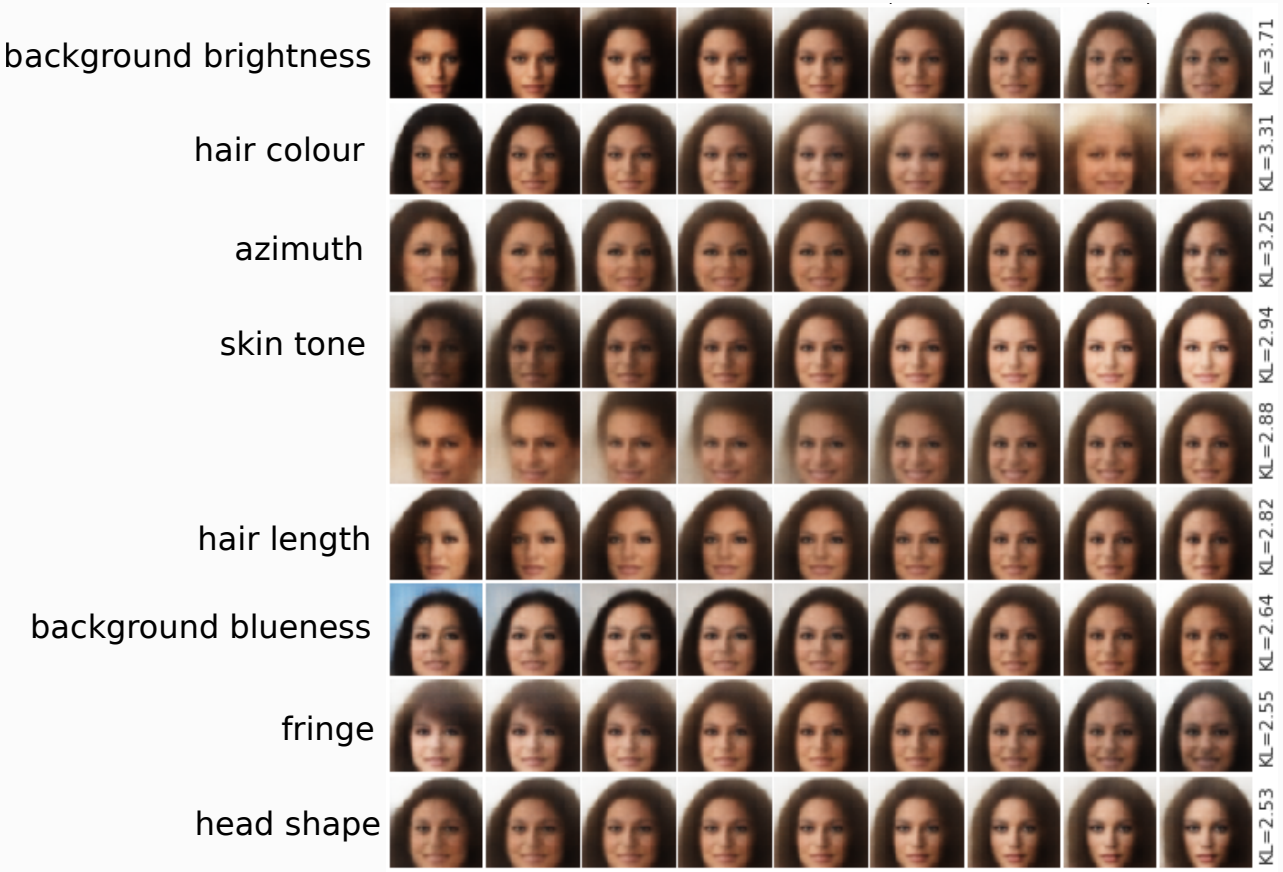
$$\arg\min_\theta \mathrm{KL}(p_{\mathcal{D}}(\boldsymbol{x}) \parallel p_\theta(\boldsymbol{x})) = \arg\max_\theta \mathbb{E}_{p_{\mathcal{D}}(\boldsymbol{x})}[\log p_\theta(\boldsymbol{x})]$$

- As this intractable, instead introduce inference model $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ and replace $\log p_\theta(\boldsymbol{x})$ with the lower bound

$$\begin{aligned}\mathcal{L}(\boldsymbol{x};\theta,\phi) &\triangleq \log p_\theta(\boldsymbol{x}) - \mathrm{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \parallel p_\theta(\boldsymbol{z}|\boldsymbol{x})) \\ &= \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - \mathrm{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \parallel p(\boldsymbol{z})).\end{aligned} \tag{1}$$

- VAE views this as a deep stochastic autoencoder, where $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ is the encoder and $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ the decoder

Disentanglement



[1] Hyunjik Kim and Andriy Mnih. "Disentangling by Factorising". In: *International Conference on Machine Learning*. 2018.

Disentanglement

- Care about the pushforward distribution of the data in latent space:

$$q_\phi(\boldsymbol{z}) = \mathbb{E}_{p_\mathcal{D}(\boldsymbol{x})}\left[q_\phi(\boldsymbol{z}|\boldsymbol{x})\right] \tag{2}$$

Disentanglement

- Care about the pushforward distribution of the data in latent space:

$$q_\phi(\boldsymbol{z}) = \mathbb{E}_{p_\mathcal{D}(\boldsymbol{x})}\left[q_\phi(\boldsymbol{z}|\boldsymbol{x})\right] \tag{2}$$

- Approaches try to learn representations where the dimensions of $\boldsymbol{z}$ are independent, i.e.

$$q_\phi(\boldsymbol{z}) = \prod_d q_\phi(z_d) \tag{3}$$

Disentanglement

- Care about the pushforward distribution of the data in latent space:

$$q_\phi(\boldsymbol{z}) = \mathbb{E}_{p_\mathcal{D}(\boldsymbol{x})}\left[q_\phi(\boldsymbol{z}|\boldsymbol{x})\right] \tag{2}$$

- Approaches try to learn representations where the dimensions of $\boldsymbol{z}$ are independent, i.e.

$$q_\phi(\boldsymbol{z}) = \prod_d q_\phi(z_d) \tag{3}$$

- Some papers require z_d to correspond to "true generative factors"

Decomposition: A Generalization of Disentanglement

More general framework for imposing *meaningful* structure on learned representations

Decomposition: A Generalization of Disentanglement

More general framework for imposing *meaningful* structure on learned representations

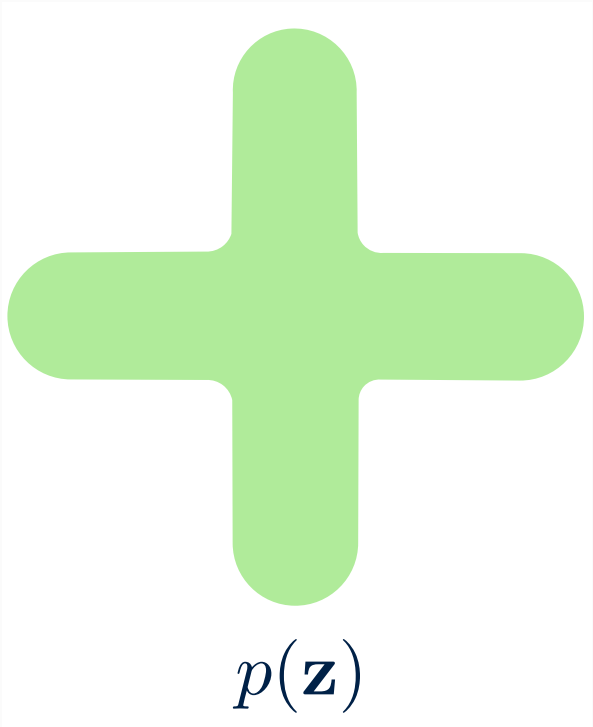
Characterise decomposition as the fulfilment of two factors:

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

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

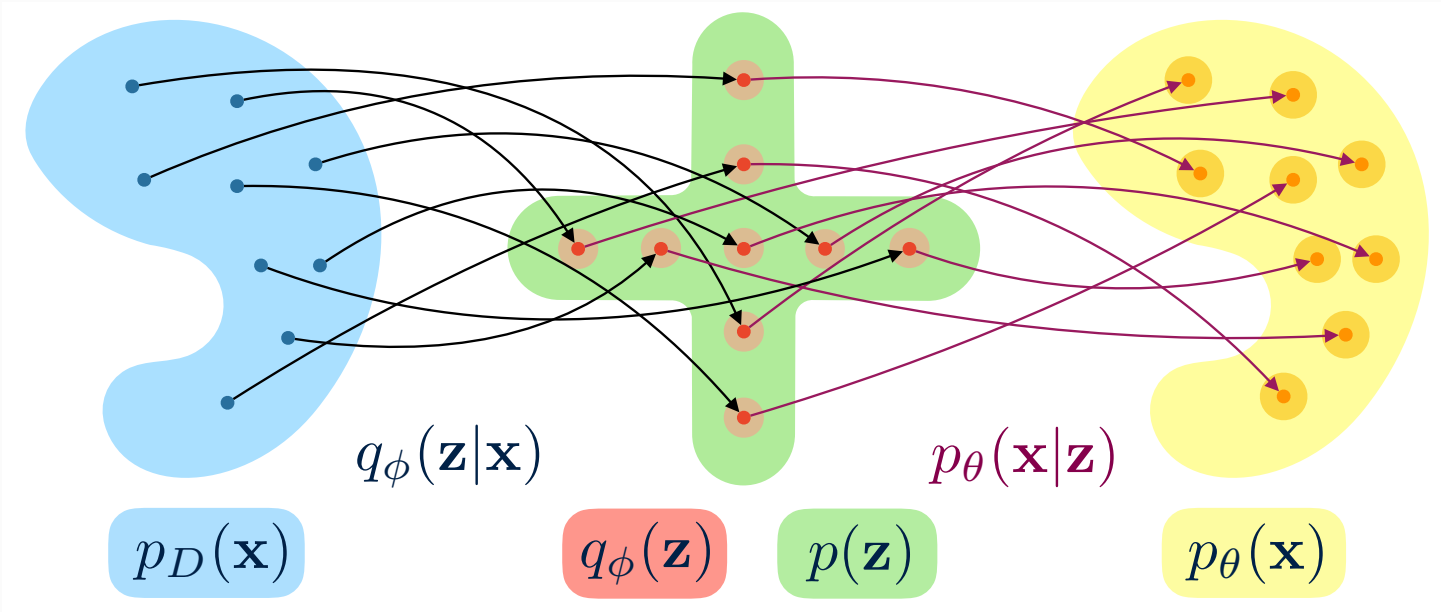
Decomposition: An Analysis

Desired Structure



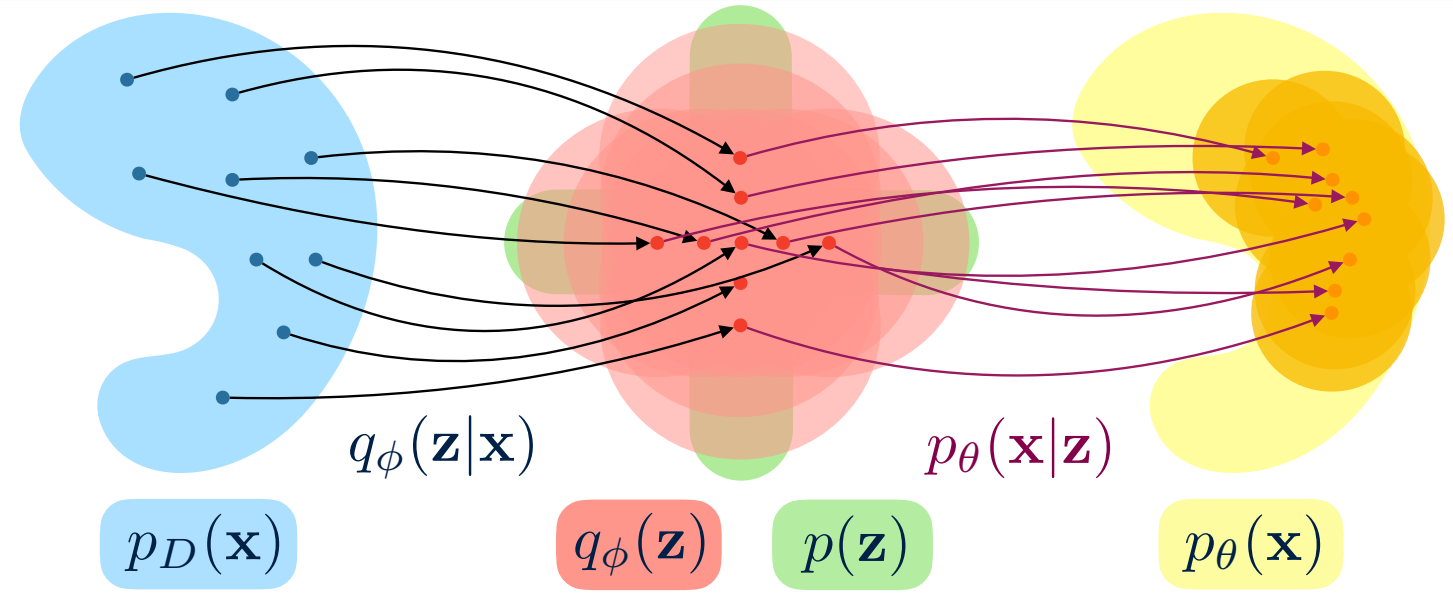
Decomposition: An Analysis

Insufficient Overlap



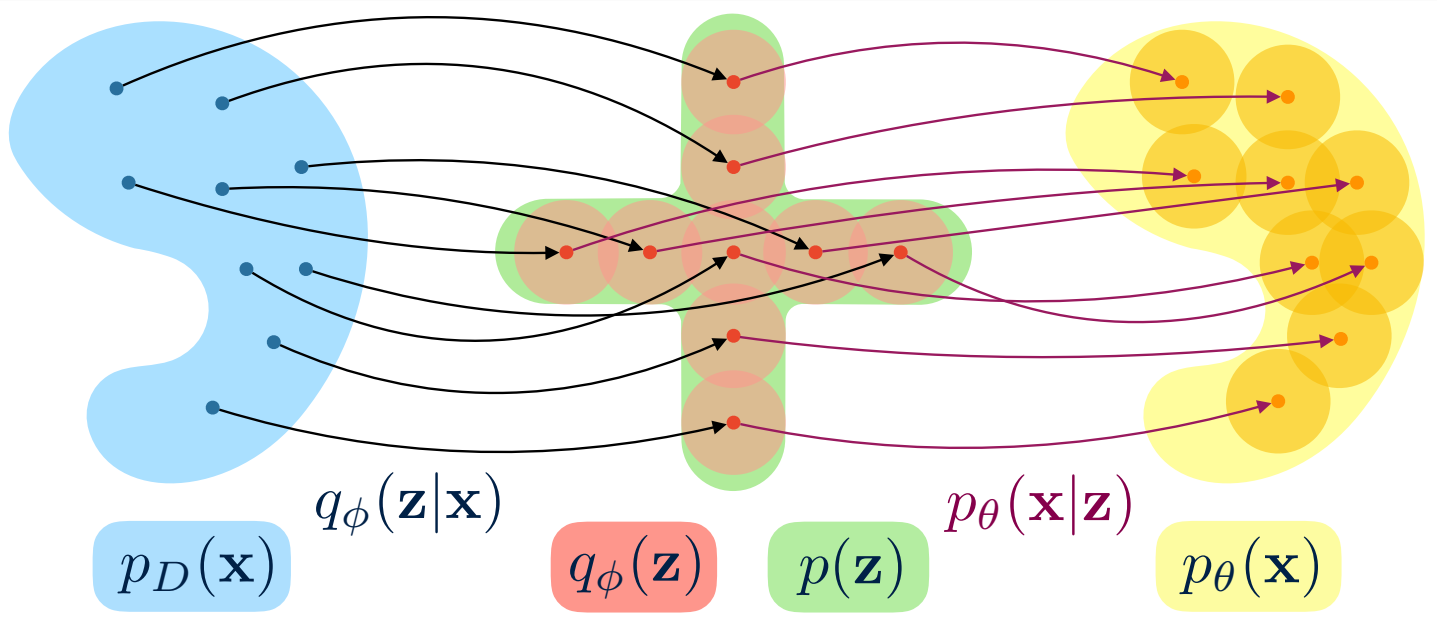
Decomposition: An Analysis

Too Much Overlap



Decomposition: An Analysis

Appropriate Overlap



Deconstructing the β-VAE

The β-VAE[2] adjusts the standard ELBO by scaling the KL term:

$$\mathcal{L}_\beta(\boldsymbol{x}) = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})] - \beta \cdot \mathrm{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})) \tag{4}$$

[2] Irina Higgins et al. "beta-VAE: Learning basic visual concepts with a constrained variational framework". In: *Proceedings of the International Conference on Learning Representations*. 2016.

Deconstructing the β-VAE — Entropy Regulariser

Theorem 1

$\mathcal{L}_\beta(\boldsymbol{x})$ can be interpreted in terms of the standard ELBO for an adjusted target $\pi_{\theta,\beta}(\boldsymbol{x},\boldsymbol{z}) \triangleq p_\theta(\boldsymbol{x} \mid \boldsymbol{z}) f_\beta(\boldsymbol{z})$ with annealed prior $f_\beta(\boldsymbol{z}) \triangleq p(\boldsymbol{z})^\beta / F_\beta$

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

Deconstructing the β-VAE — Entropy Regulariser

Theorem 1

$\mathcal{L}_\beta(\boldsymbol{x})$ can be interpreted in terms of the standard ELBO for an adjusted target $\pi_{\theta,\beta}(\boldsymbol{x},\boldsymbol{z}) \triangleq p_\theta(\boldsymbol{x} \mid \boldsymbol{z}) f_\beta(\boldsymbol{z})$ with annealed prior $f_\beta(\boldsymbol{z}) \triangleq p(\boldsymbol{z})^\beta / F_\beta$

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

Implications

β-VAE disentangles largely by controlling the level of overlap

Deconstructing the β-VAE — Entropy Regulariser

Theorem 1

$\mathcal{L}_\beta(\boldsymbol{x})$ can be interpreted in terms of the standard ELBO for an adjusted target $\pi_{\theta,\beta}(\boldsymbol{x},\boldsymbol{z}) \triangleq p_\theta(\boldsymbol{x} \mid \boldsymbol{z}) f_\beta(\boldsymbol{z})$ with annealed prior $f_\beta(\boldsymbol{z}) \triangleq p(\boldsymbol{z})^\beta / F_\beta$

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

Implications

β-VAE disentangles largely by controlling the level of overlap

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

Deconstructing the β-VAE — Rotational Invariance

Theorem 2

If $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; 0, \sigma I)$ *and* $q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{z}; \mu_\phi(\boldsymbol{x}), S_\phi(\boldsymbol{x})\right)$*, then for all rotation matrices* R*,*

$$\mathcal{L}_\beta(\boldsymbol{x}; \theta, \phi) = \mathcal{L}_\beta(\boldsymbol{x}; \theta^\dagger(R), \phi^\dagger(R))$$

where $\theta^\dagger(R)$ *and* $\phi^\dagger(R)$ *are transformed networks such that*

$$\begin{aligned} p_{\theta^\dagger}(\boldsymbol{x} \mid \boldsymbol{z}) &= p_\theta\left(\boldsymbol{x} \mid R^T \boldsymbol{z}\right), \\ q_{\phi^\dagger}(\boldsymbol{z}|\boldsymbol{x}) &= \mathcal{N}\left(\boldsymbol{z}; R\mu_\phi(\boldsymbol{x}), RS_\phi(\boldsymbol{x})R^T\right). \end{aligned}$$

Deconstructing the β-VAE — Rotational Invariance

Theorem 2

If $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; 0, \sigma I)$ *and* $q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) = \mathcal{N}\left(\boldsymbol{z}; \mu_\phi(\boldsymbol{x}), S_\phi(\boldsymbol{x})\right)$*, then for all rotation matrices* R*,*

$$\mathcal{L}_\beta(\boldsymbol{x}; \theta, \phi) = \mathcal{L}_\beta(\boldsymbol{x}; \theta^\dagger(R), \phi^\dagger(R))$$

where $\theta^\dagger(R)$ *and* $\phi^\dagger(R)$ *are transformed networks such that*

$$\begin{aligned} p_{\theta^\dagger}(\boldsymbol{x} \mid \boldsymbol{z}) &= p_\theta\left(\boldsymbol{x} \mid R^T \boldsymbol{z}\right), \\ q_{\phi^\dagger}(\boldsymbol{z}|\boldsymbol{x}) &= \mathcal{N}\left(\boldsymbol{z}; R\mu_\phi(\boldsymbol{x}), RS_\phi(\boldsymbol{x})R^T\right). \end{aligned}$$

Implications

The β-VAE does not directly encourage meaningful representations when $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; 0, \sigma I)$

Decomposition: Objective

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

$$- \beta \cdot \mathrm{KL}(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \,\|\, p(\boldsymbol{z})) \qquad \text{Control level of overlap}$$

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

Decomposition: Generalising Disentanglement

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

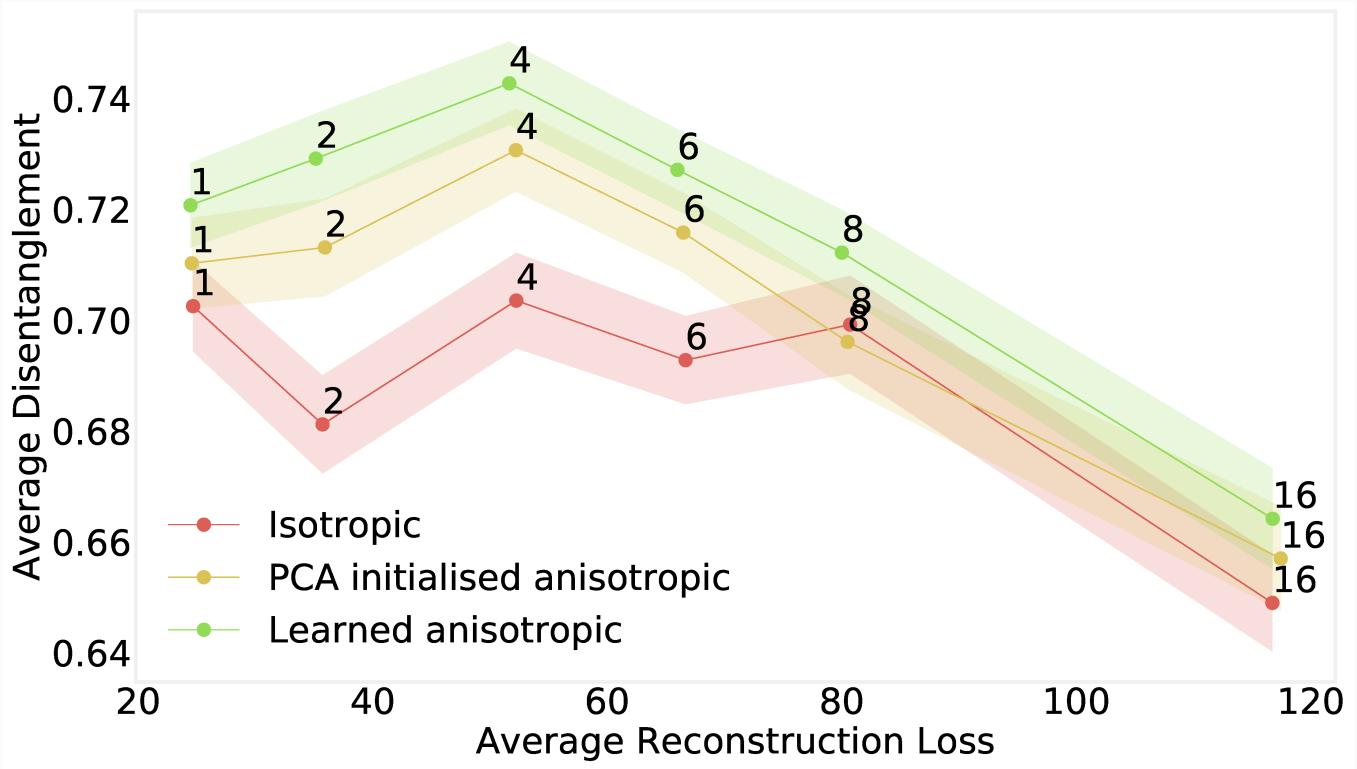


Figure 1: β-VAE trained on *2D Shapes*[3] computing disentanglement[4].

[3] Loic Matthey et al. *dSprites: Disentanglement testing Sprites dataset.* https://github.com/deepmind/dsprites-dataset/. 2017.

[4] Kim and Mnih, "Disentangling by Factorising".

Decomposition: Generalising Disentanglement

Clustering: $p(\boldsymbol{z}) = \sum_{k=1}^{4} \rho_k \cdot \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k)$

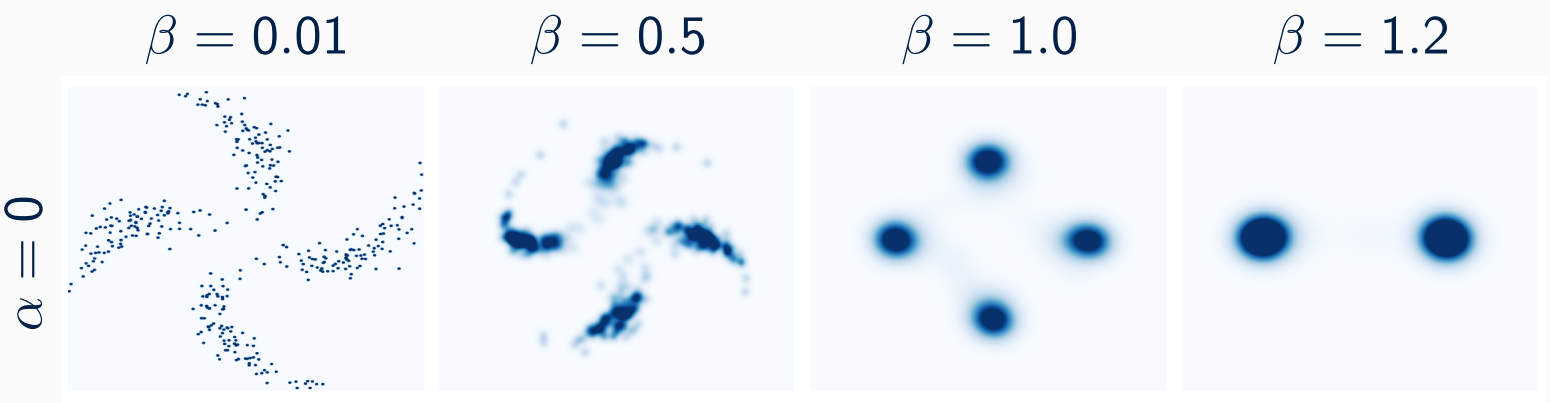


Figure 2: Density of aggregate posterior $q_\phi(\boldsymbol{z})$ with different α, β for the pinwheel dataset.[5]

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

Decomposition: Generalising Disentanglement

Clustering: $p(\boldsymbol{z}) = \sum_{k=1}^{4} \rho_k \cdot \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k)$

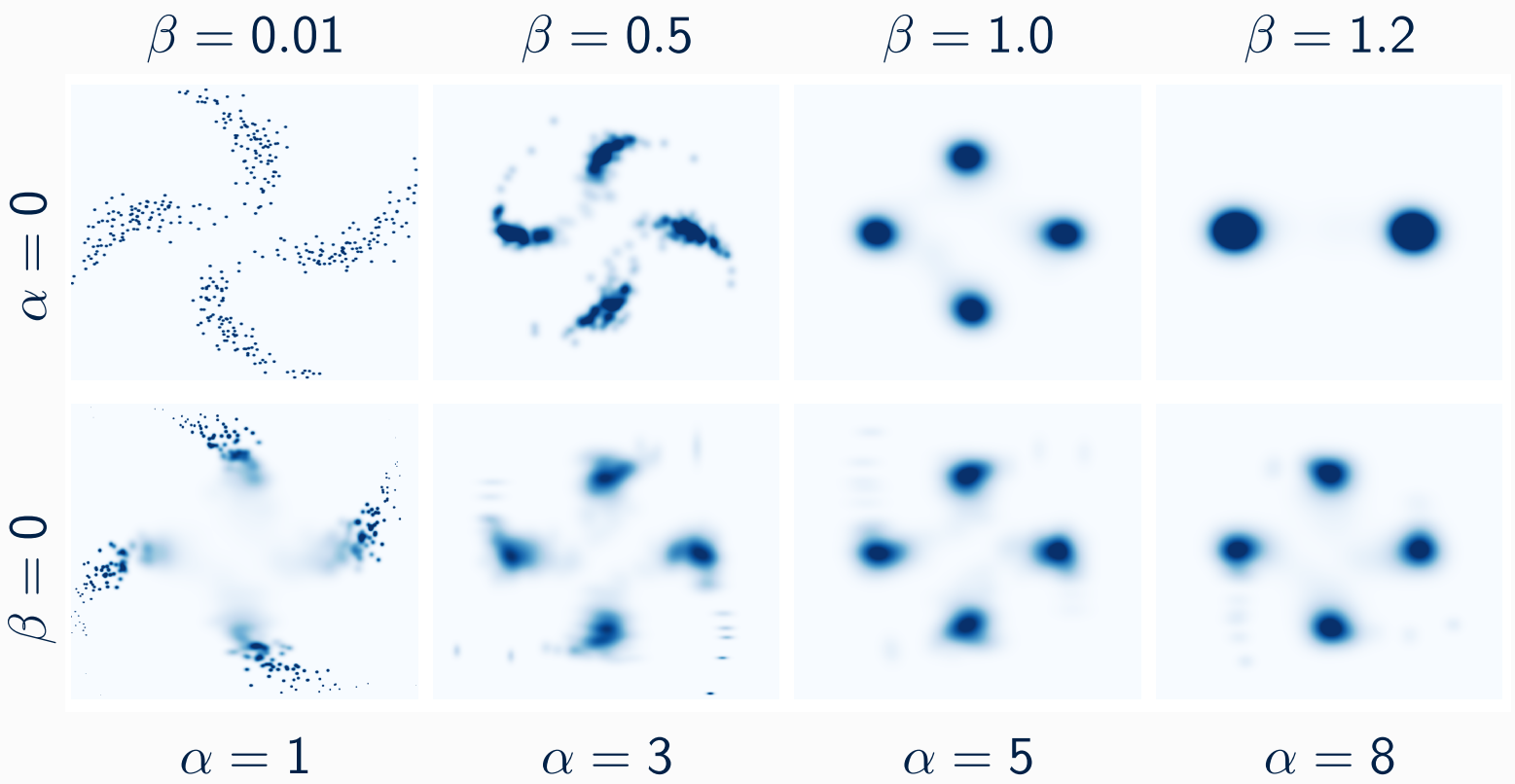


Figure 2: Density of aggregate posterior $q_\phi(\boldsymbol{z})$ with different α, β for the pinwheel dataset.[5]

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

Decomposition: Generalising Disentanglement

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

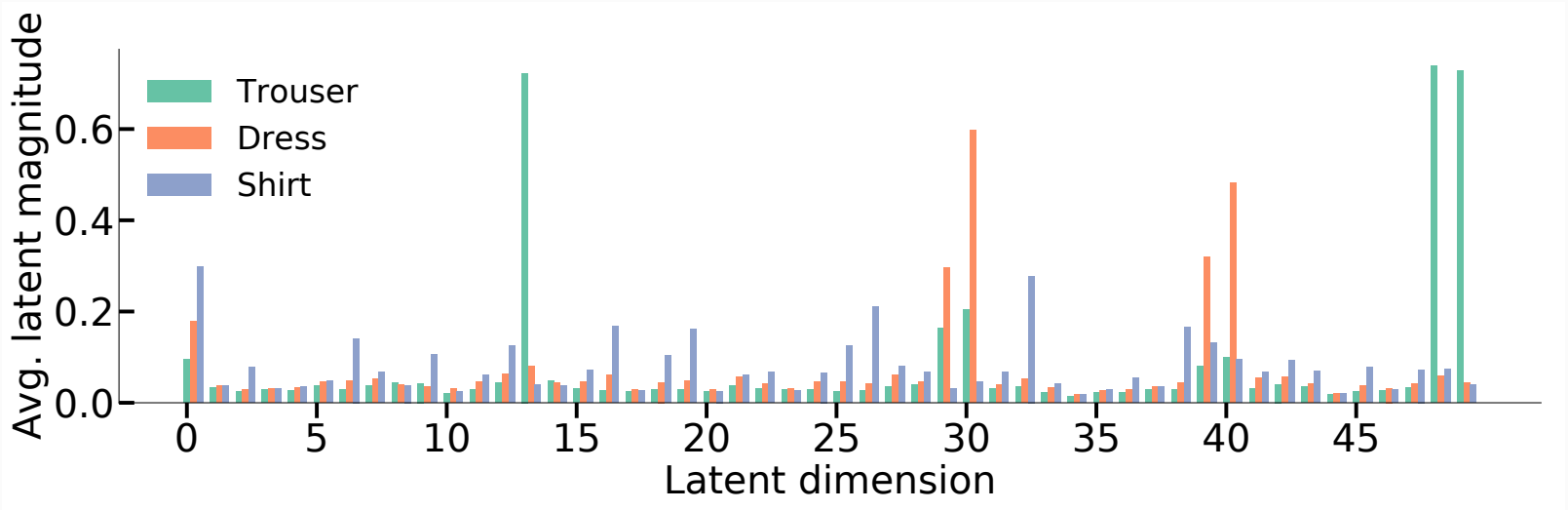
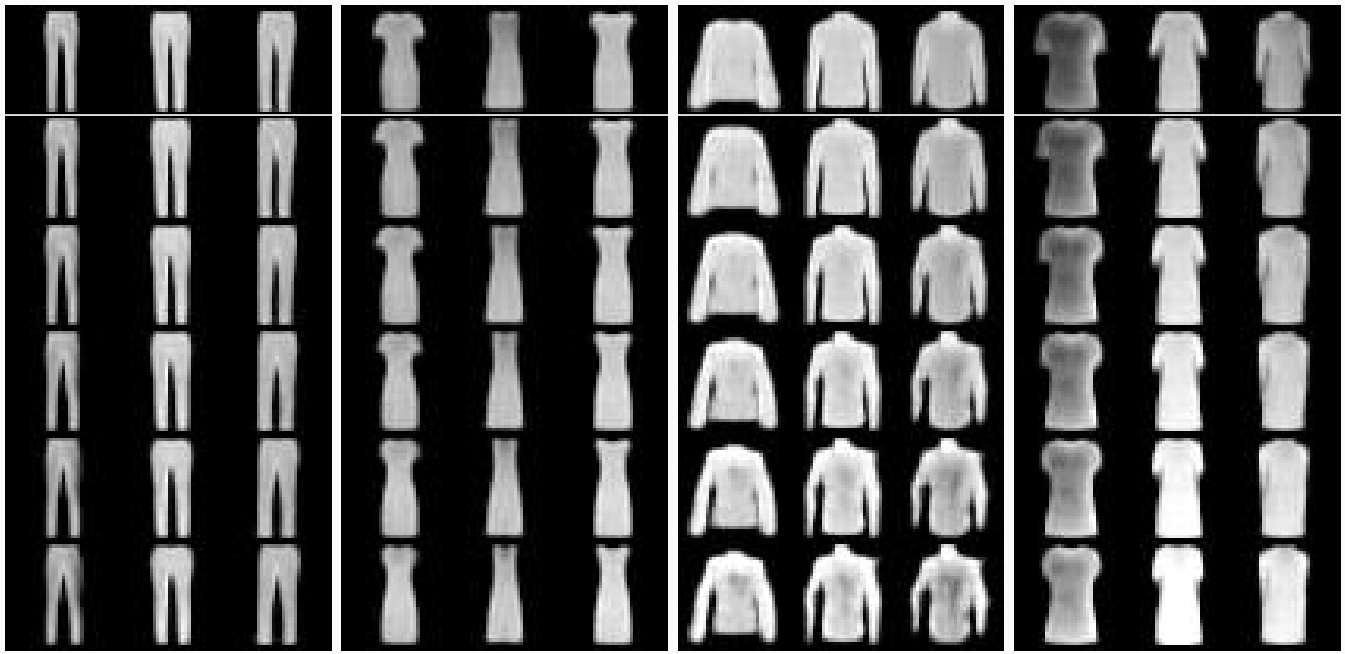


Figure 3: Sparsity of learnt representations for the *Fashion-MNIST*[6] dataset.

[6] Han Xiao, Kashif Rasul, and Roland Vollgraf. *Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms*. Aug. 28, 2017. arXiv: cs.LG/1708.07747 [cs.LG].

Decomposition: Generalising Disentanglement

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



(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[6].

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

Decomposition: Generalising Disentanglement

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

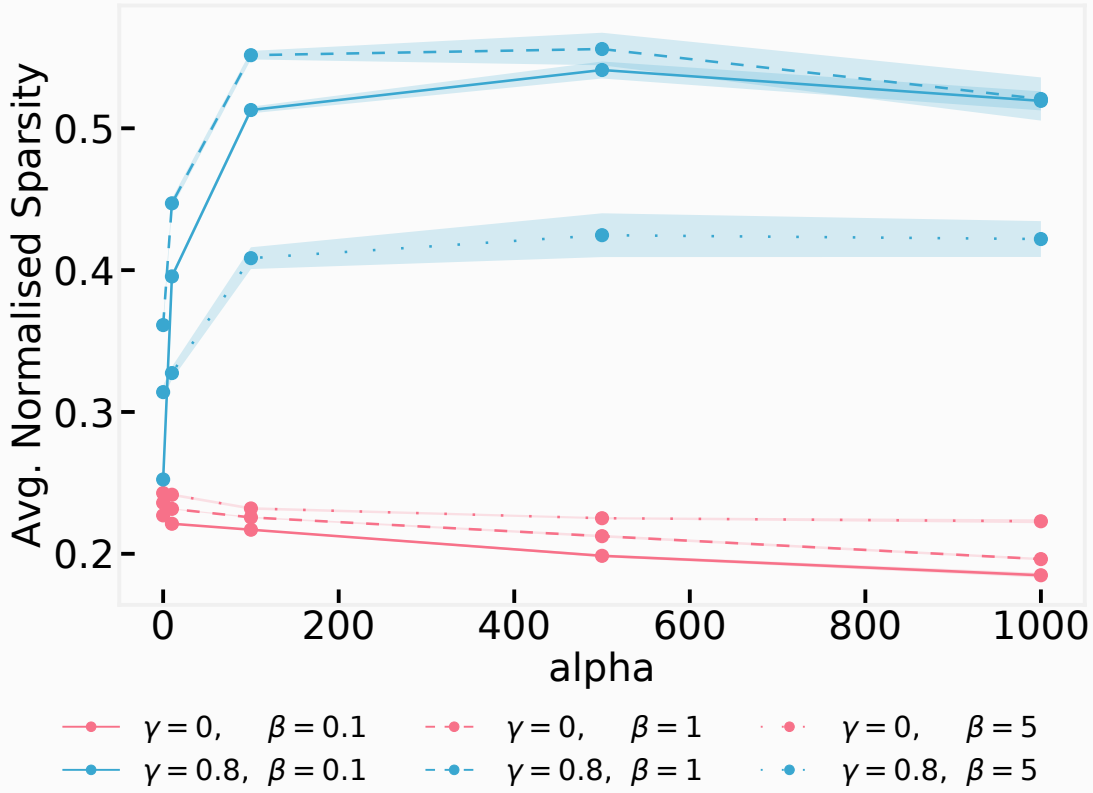


Figure 3: Sparsity vs regularisation strength α (higher better)[6].

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

Recap

We propose and develop:

Recap

We propose and develop:

- Decomposition: a generalisation of disentanglement involving:

Recap

We propose and develop:

- Decomposition: a generalisation of disentanglement involving:
  (a) match between $q_\phi(\boldsymbol{z})$ and $p(\boldsymbol{z})$

Recap

We propose and develop:

- Decomposition: a generalisation of disentanglement involving:
  (a) match between $q_\phi(\boldsymbol{z})$ and $p(\boldsymbol{z})$
  (b) appropriate overlap of latent encodings

Recap

We propose and develop:

- Decomposition: a generalisation of disentanglement involving:
  (a) match between $q_\phi(\boldsymbol{z})$ and $p(\boldsymbol{z})$
  (b) appropriate overlap of latent encodings
- A theoretical analysis of the β-VAE objective showing it primarily only contributes to overlap.

Recap

We propose and develop:

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

Recap

We propose and develop:

- Decomposition: a generalisation of disentanglement involving:
  (a) match between $q_\phi(\boldsymbol{z})$ and $p(\boldsymbol{z})$
  (b) appropriate overlap of latent encodings
- 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

Discussion Points

How Do We Characterize Overlap?

Discussion Points

How Do We Characterize Overlap?

- Closely linked to the mutual information $I(\boldsymbol{x}; \boldsymbol{z})$

Discussion Points

How Do We Characterize Overlap?

- Closely linked to the mutual information $I(\boldsymbol{x}; \boldsymbol{z})$
- But two can noticeably vary, particularly for complicated encoders

Discussion Points

How Do We Characterize Overlap?

- Closely linked to the mutual information $I(\boldsymbol{x}; \boldsymbol{z})$
- But two can noticeably vary, particularly for complicated encoders
- Can achieve any $I(\boldsymbol{x}; \boldsymbol{z})$ with negligible overlap if

$$q_\phi(\boldsymbol{z}|\boldsymbol{x}) = \lambda \cdot \text{Uniform}\left(\|\mu_\phi(\boldsymbol{x}) - \boldsymbol{z}\|_2 < \epsilon\right) + (1 - \lambda) \cdot p(\boldsymbol{z}) \qquad (5)$$

Discussion Points

How Do We Characterize Overlap?

- Closely linked to the mutual information $I(\boldsymbol{x}; \boldsymbol{z})$
- But two can noticeably vary, particularly for complicated encoders
- Can achieve any $I(\boldsymbol{x}; \boldsymbol{z})$ with negligible overlap if

$$q_\phi(\boldsymbol{z}|\boldsymbol{x}) = \lambda \cdot \text{Uniform}\left(\|\mu_\phi(\boldsymbol{x}) - \boldsymbol{z}\|_2 < \epsilon\right) + (1 - \lambda) \cdot p(\boldsymbol{z}) \tag{5}$$

- $I(\boldsymbol{x}; \boldsymbol{z})$ does not distinguish between large overlap with a small number of datapoints and small overlap with a large number

Discussion Points

How Do We Characterize Overlap?

- Closely linked to the mutual information $I(\boldsymbol{x}; \boldsymbol{z})$
- But two can noticeably vary, particularly for complicated encoders
- Can achieve any $I(\boldsymbol{x}; \boldsymbol{z})$ with negligible overlap if

$$q_\phi(\boldsymbol{z}|\boldsymbol{x}) = \lambda \cdot \text{Uniform}\left(\|\mu_\phi(\boldsymbol{x}) - \boldsymbol{z}\|_2 < \epsilon\right) + (1 - \lambda) \cdot p(\boldsymbol{z}) \tag{5}$$

- $I(\boldsymbol{x}; \boldsymbol{z})$ does not distinguish between large overlap with a small number of datapoints and small overlap with a large number
- It can also fail to account for *locality* in the latent space

Discussion Points

Can VAEs Uncover True Generative Factors?

[7] Francesco Locatello et al. "Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations". In: *International Conference on Machine Learning* (2019).

Discussion Points

Can VAEs Uncover True Generative Factors?

- Empirically this is well known to be very difficult

[7] Locatello et al., "Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations".

Discussion Points

Can VAEs Uncover True Generative Factors?

- Empirically this is well known to be very difficult
- Locatello et al. (2019)[7] suggest unsupervised disentanglement is impossible without inductive biases due to equivalence classes in the generative model

[7] Locatello et al., "Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations".

Discussion Points

Can VAEs Uncover True Generative Factors?

- Empirically this is well known to be very difficult
- Locatello et al. (2019)[7] suggest unsupervised disentanglement is impossible without inductive biases due to equivalence classes in the generative model
- This fails to account for the stochasticity of the encoder so do not directly apply to VAEs

[7] Locatello et al., "Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations".

Discussion Points

Can VAEs Uncover True Generative Factors?

- Empirically this is well known to be very difficult
- Locatello et al. (2019)[7] suggest unsupervised disentanglement is impossible without inductive biases due to equivalence classes in the generative model
- This fails to account for the stochasticity of the encoder so do not directly apply to VAEs
- When, and to what extent, can we expect learned representations to mimic the true generative process?

[7] Locatello et al., "Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations".

Discussion Points

Can VAEs Uncover True Generative Factors?

- Empirically this is well known to be very difficult
- Locatello et al. (2019)[7] suggest unsupervised disentanglement is impossible without inductive biases due to equivalence classes in the generative model
- This fails to account for the stochasticity of the encoder so do not directly apply to VAEs
- When, and to what extent, can we expect learned representations to mimic the true generative process?
- Should disentanglement/decomposition metrics include any notion of a true generative process?

[7] Locatello et al., "Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations".

Emile Mathieu

Tom Rainforth

N. Siddharth

Yee Whye Teh

Code

Paper

iffsid/disentangling-disentanglement

ICML 2019 arXiv:1812.02833