Deep Generative Models
Lecture 11

Roman Isachenko

Ozon Masters

2021

Disentangled representations

Unsupervised representation learning

Learning an interpretable factorised representation of the independent data generative factors of the world without supervision.

Disentanglement informal definition

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.

Example

Model trained on a dataset of 3D objects might learn independent latent units sensitive to single independent data generative factors, such as object identity, position, scale, lighting or colour.

Higgins I. et al. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework, 2017

β-VAE, 2017

Generative process

- $p(\mathbf{x}|\mathbf{v}, \mathbf{w}) = \text{Sim}(\mathbf{v}, \mathbf{w})$ – true world simulator;
- $\mathbf{v}$ – conditionally independent factors: $p(\mathbf{v}|\mathbf{x}) = \prod_{j=1}^{d} p(v_j|\mathbf{x})$;
- $\mathbf{w}$ – conditionally dependent factors.

Goal

Construct an unsupervised deep generative model

$$p(\mathbf{x}|\mathbf{z}) \approx p(\mathbf{x}|\mathbf{v}, \mathbf{w}).$$

- Ensure that the inferred latent factors $q(\mathbf{z}|\mathbf{x})$ capture the factors $\mathbf{v}$ in a disentangled manner.
- The conditionally dependent factors $\mathbf{w}$ can remain entangled in a separate subset of $\mathbf{z}$ that is not used for representing $\mathbf{v}$.

Higgins I. et al. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework, 2017

InfoGAN

GAN objective

$$\min_G \max_D V(G, D)$$

$$V(G, D) = \mathbb{E}_{\pi(\mathbf{x})} \log D(\mathbf{x}) + \mathbb{E}_{p(\mathbf{z})} \log(1 - D(G(\mathbf{z})))$$

Latent vector $\mathbf{z}$ is not imposed to be disentangled.
InfoGAN decomposes input vector:

- $\mathbf{z}$ – incompressible noise;
- $\mathbf{c}$ – structured latent code $p(\mathbf{c}) = \prod_{j=1}^{d} p(c_j)$.

Information-theoretic regularization

$$\max I(\mathbf{c}, G(\mathbf{z}, \mathbf{c}))$$

Information in the latent code $\mathbf{c}$ should not be lost in the generation process.

Chen X. et al. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets, 2016

InfoGAN

Objective

$$\min_G \max_D V(G, D) - \lambda I(\mathbf{c}, G(\mathbf{z}, \mathbf{c}))$$

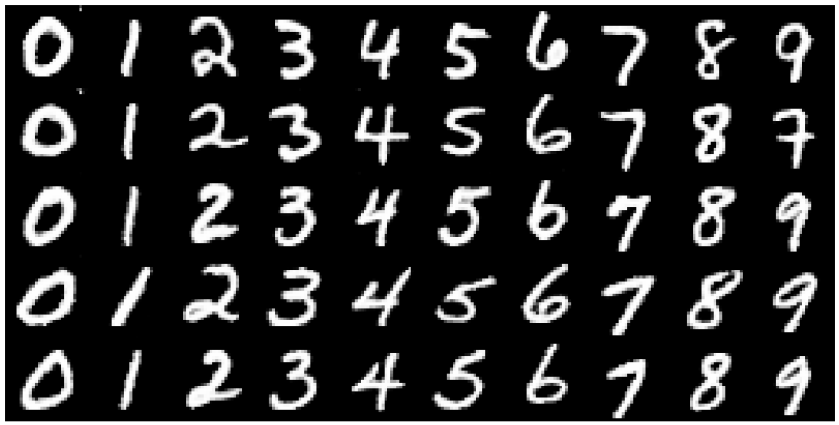
Variational Information Maximization

$$\begin{aligned} I(\mathbf{c}, G(\mathbf{z}, \mathbf{c})) &= H(\mathbf{c}) - H(\mathbf{c}|G(\mathbf{z}, \mathbf{c})) = \\ &= H(\mathbf{c}) + \mathbb{E}_{\mathbf{x}\sim G(\mathbf{z},\mathbf{c})}\left[\mathbb{E}_{\mathbf{c}'\sim p(\mathbf{c}|\mathbf{x})} \log p(\mathbf{c}'|\mathbf{x})\right] = \\ &= H(\mathbf{c}) + \mathbb{E}_{\mathbf{x}\sim G(\mathbf{z},\mathbf{c})} KL(p(\mathbf{c}'|\mathbf{x})||q(\mathbf{z}'|\mathbf{x})) + \\ &\quad + \mathbb{E}_{\mathbf{x}\sim G(\mathbf{z},\mathbf{c})}\mathbb{E}_{\mathbf{c}'\sim p(\mathbf{c}|\mathbf{x})} \log q(\mathbf{c}'|\mathbf{x}) \geq \\ &\geq H(\mathbf{c}) + \mathbb{E}_{\mathbf{x}\sim G(\mathbf{z},\mathbf{c})}\mathbb{E}_{\mathbf{c}'\sim p(\mathbf{c}|\mathbf{x})} \log q(\mathbf{c}'|\mathbf{x}) = \\ &\quad H(\mathbf{c}) + \mathbb{E}_{\mathbf{c}\sim p(\mathbf{c})}\mathbb{E}_{\mathbf{x}\sim G(\mathbf{z},\mathbf{c})} \log q(\mathbf{c}|\mathbf{x}) \end{aligned}$$

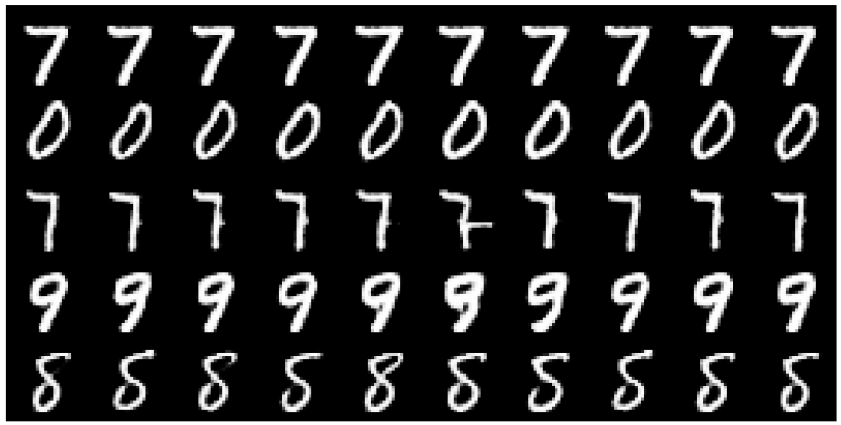
Chen X. et al. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets, 2016

InfoGAN

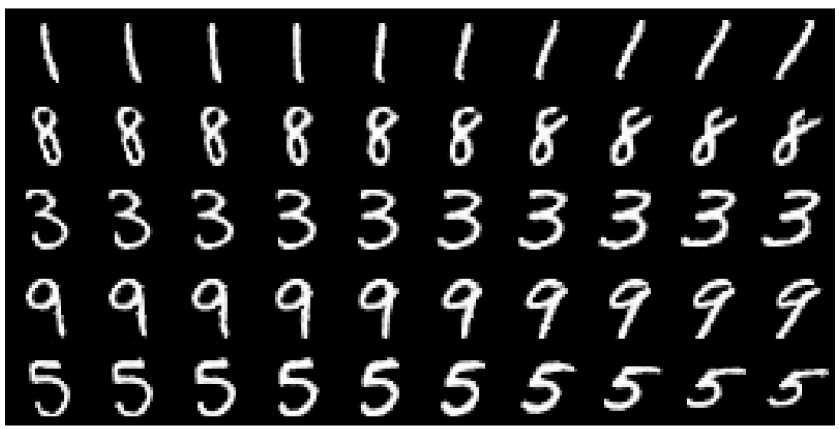
Latent codes on MNIST



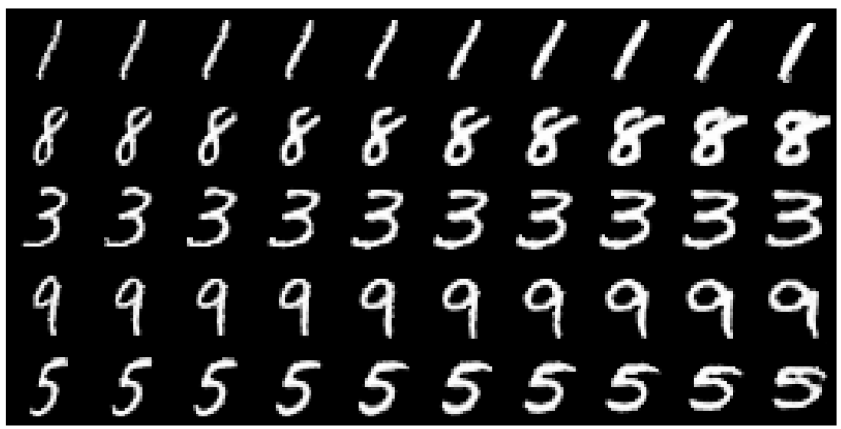
(a) Varying c_1 on InfoGAN (Digit type)



(b) Varying c_1 on regular GAN (No clear meaning)



(c) Varying c_2 from -2 to 2 on InfoGAN (Rotation)



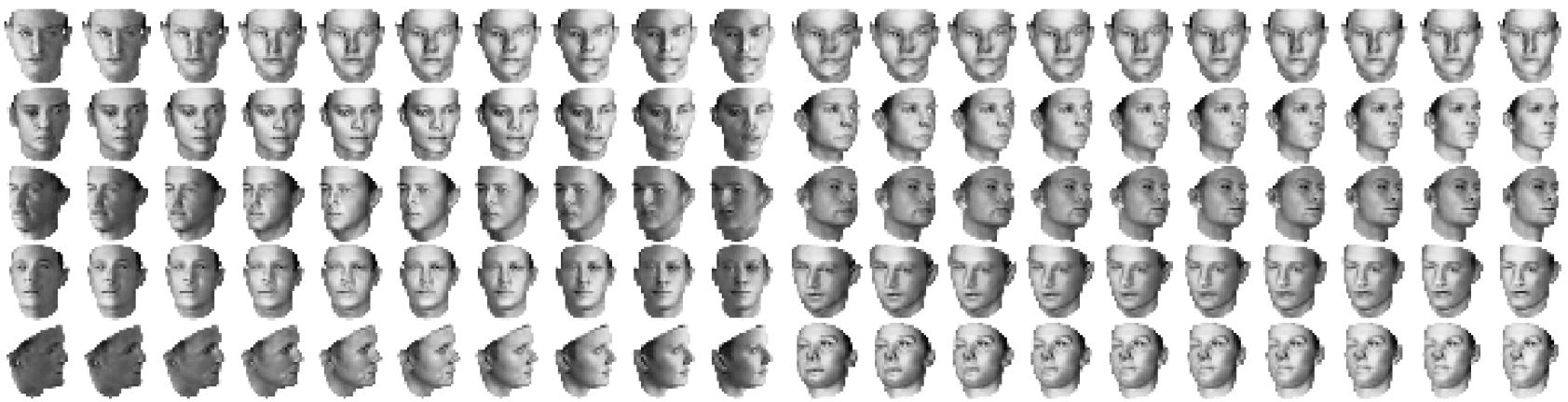
(d) Varying c_3 from -2 to 2 on InfoGAN (Width)

Chen X. et al. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets, 2016

InfoGAN

Latent codes on 3D Faces

(a) Azimuth (pose) (b) Elevation



(c) Lighting (d) Wide or Narrow

Chen X. et al. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets, 2016

β-VAE

Constrained optimization

$$\max_{q,\boldsymbol{\theta}} \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \log p(\mathbf{x}|\mathbf{z},\boldsymbol{\theta}), \quad \text{subject to } KL(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) < \epsilon.$$

Objective

$$\mathcal{L}(q,\boldsymbol{\theta},\beta) = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \log p(\mathbf{x}|\mathbf{z},\boldsymbol{\theta}) - \beta \cdot KL(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})).$$

What do we get at $\beta = 1$?

Hypothesis

To learn disentangled representations of the conditionally independent factors $\mathbf{v}$, it is important to set a stronger constraint on the latent bottleneck: $\beta > 1$.

Note: It leads to poorer reconstructions and a loss of high frequency details when passing through a constrained latent bottleneck.

Higgins I. et al. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework, 2017

β-VAE

Disentangling metric

1. Generate two sets of objects

$$\mathbf{x}_{li} \sim \text{Sim}(\mathbf{v}_{li}, \mathbf{w}_{li}); \quad \mathbf{x}_{lj} \sim \text{Sim}(\mathbf{v}_{lj}, \mathbf{w}_{lj}); \quad y_{ij} \sim U[1, d].$$

$$\mathbf{v}_{li} \sim p(\mathbf{v}); \quad \mathbf{v}_{lj} \sim p(\mathbf{v})\,([v_{li}]_y = [v_{lj}]_y); \quad \mathbf{w}_{li}, \mathbf{w}_{lj} \sim p(\mathbf{w}).$$

2. Find representations

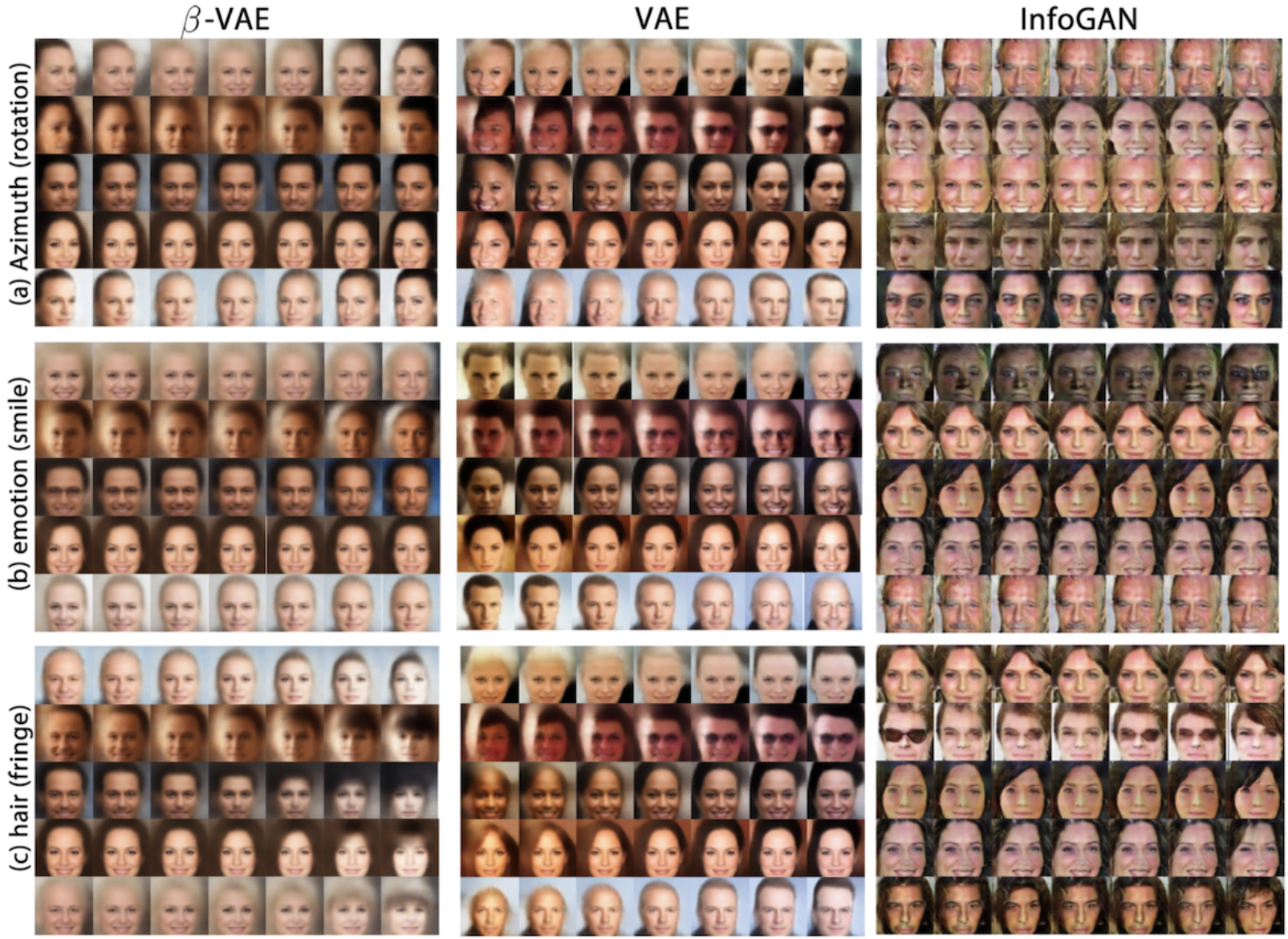
$$q(\mathbf{z}|\mathbf{x}) = \mathcal{N}\left(\mu(\mathbf{x})|\sigma^2(\mathbf{x})\right); \quad \mathbf{z}_{li} = \mu(\mathbf{x}_{li}); \quad \mathbf{z}_{lj} = \mu(\mathbf{x}_{lj}).$$

3. Use accuracy of classifier $p(y|\mathbf{z}_{\text{diff}})$ with a low VC-dimension as metric of disentanglement

$$\mathbf{z}_{\text{diff}} = \frac{1}{L}\sum_{l=1}^{L} |\mathbf{z}_{li} - \mathbf{z}_{lj}|.$$

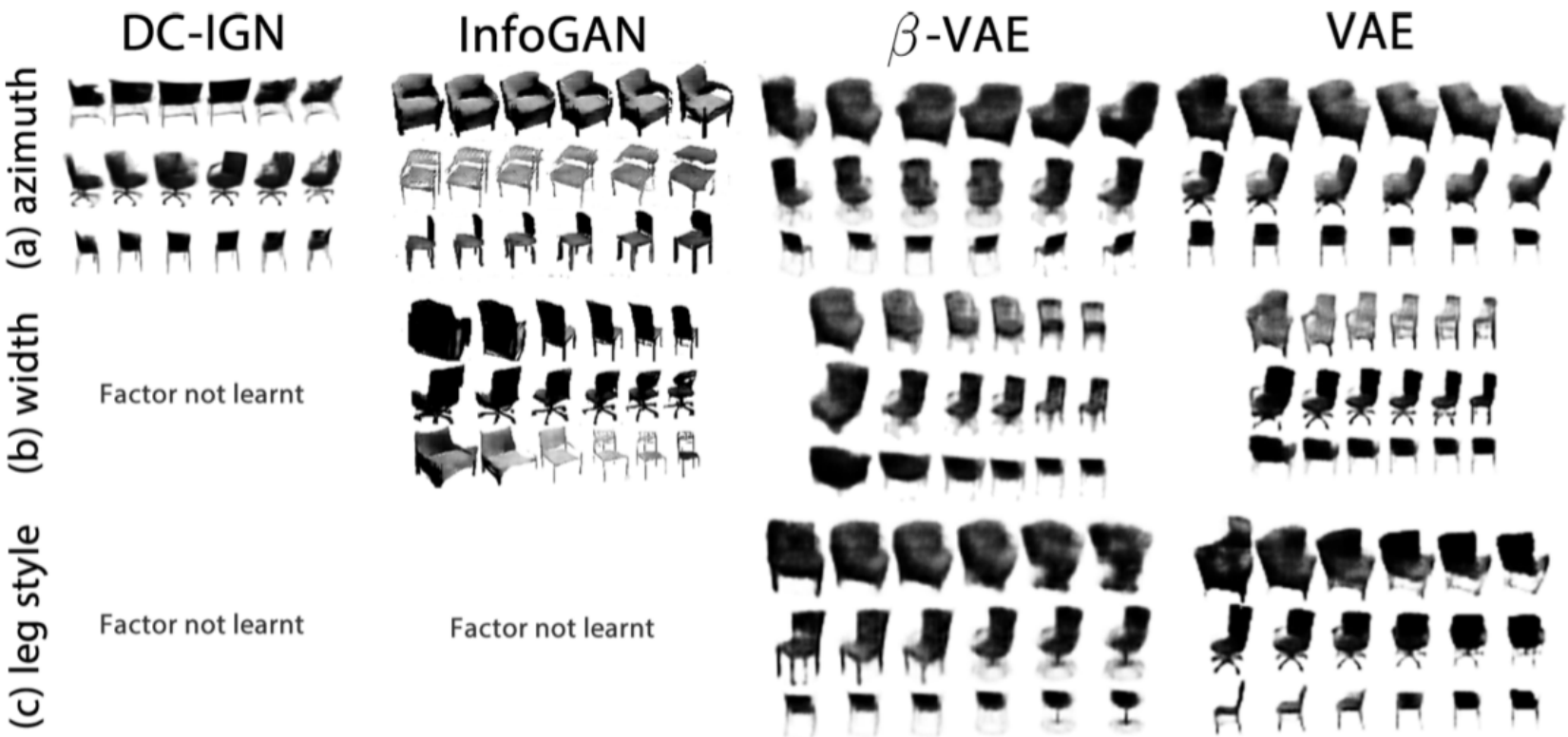
Higgins I. et al. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework, 2017

β-VAE



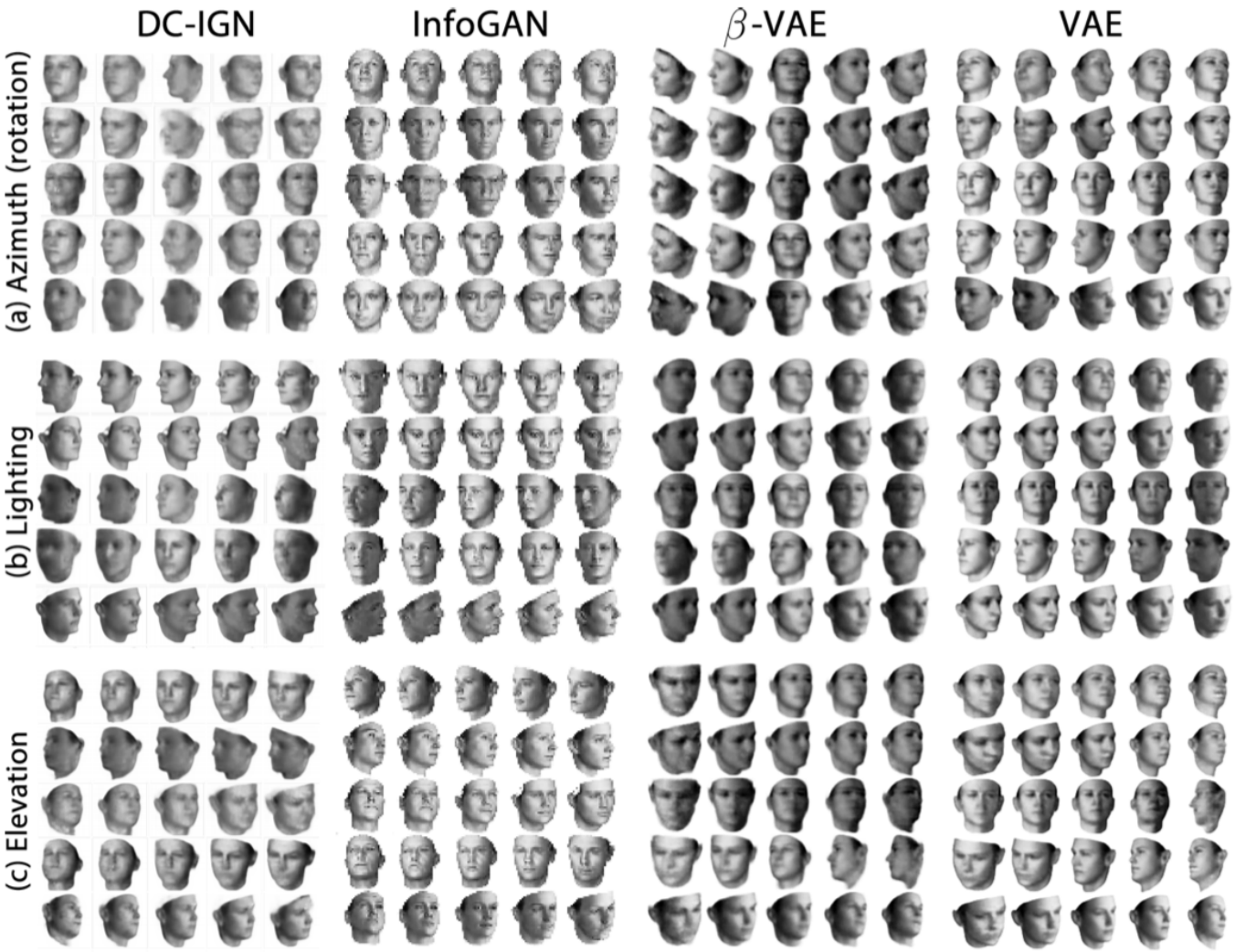
Higgins I. et al. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework, 2017

β-VAE



Higgins I. et al. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework, 2017

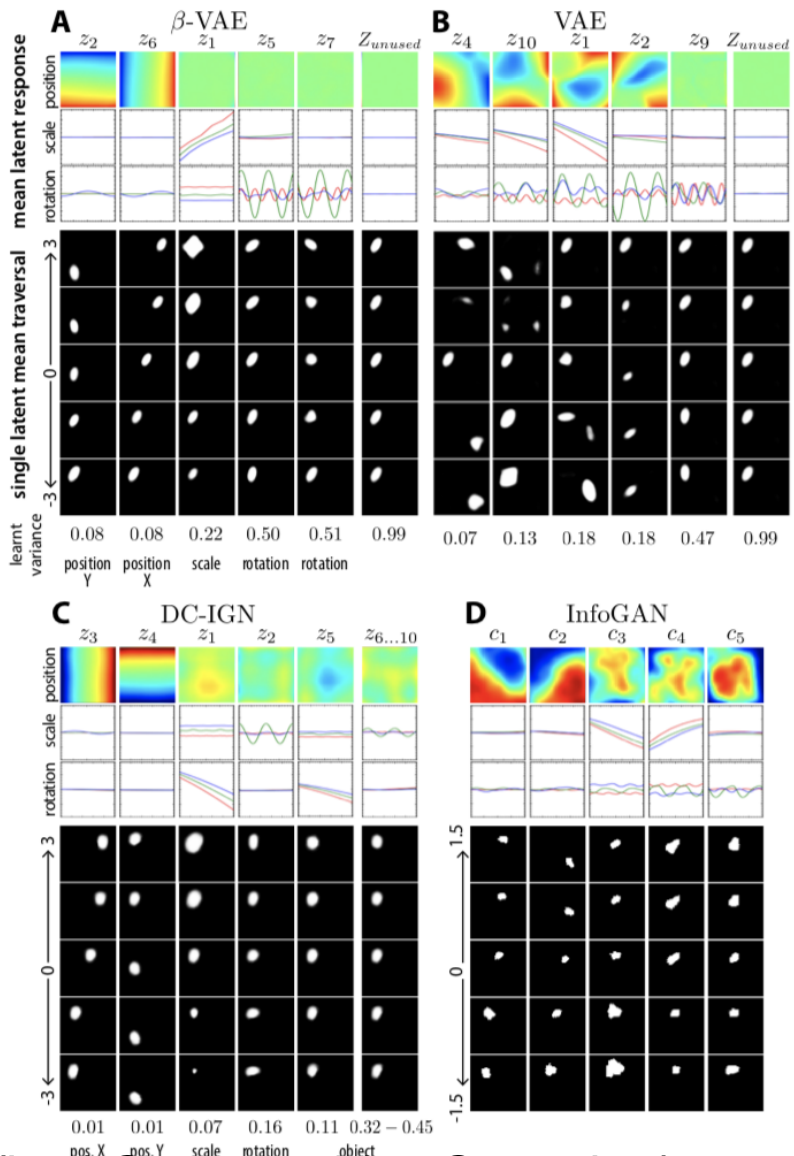
β-VAE



Higgins I. et al. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework, 2017

β-VAE

Model	Disentanglement metric score
Ground truth	*100%*
Raw pixels	$45.75 \pm 0.8\%$
PCA	$84.9 \pm 0.4\%$
ICA	$42.03 \pm 10.6\%$
DC-IGN	$\mathbf{99.3 \pm 0.1\%}$%
InfoGAN	$73.5 \pm 0.9\%$
VAE untrained	$44.14 \pm 2.5\%$
VAE	$61.58 \pm 0.5\%$
β-VAE	$\mathbf{99.23 \pm 0.1\%}$

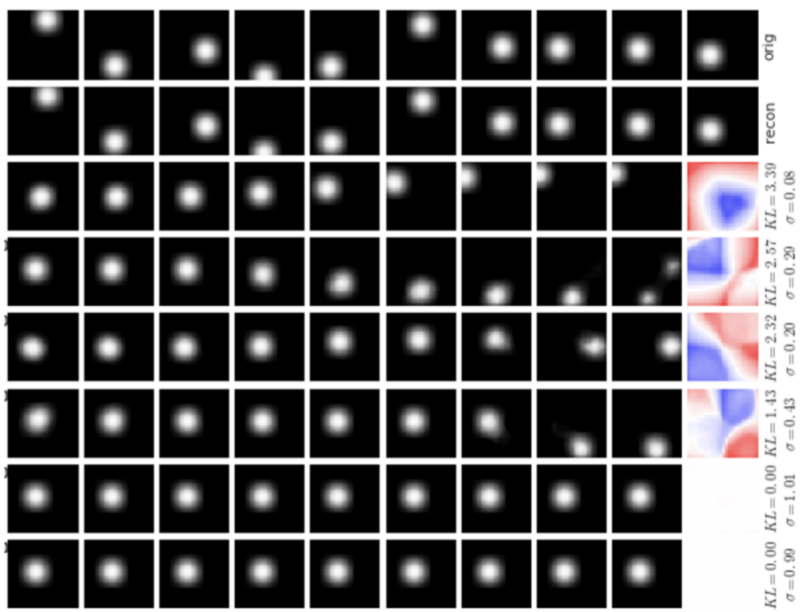


Higgins I. et al. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework, 2017

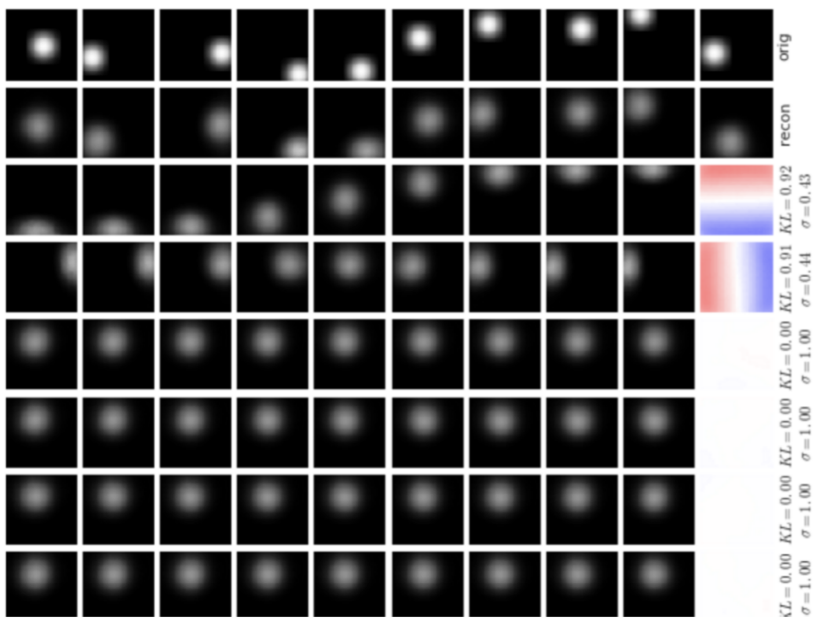
β-VAE

- **Top row:** original images.
- **Second row:** the corresponding reconstructions.
- **Remaining rows:** latent traversals ordered by KL divergence with the prior.
- **Heatmaps:** latent activations for each 2D position.

β = 1



β = 150

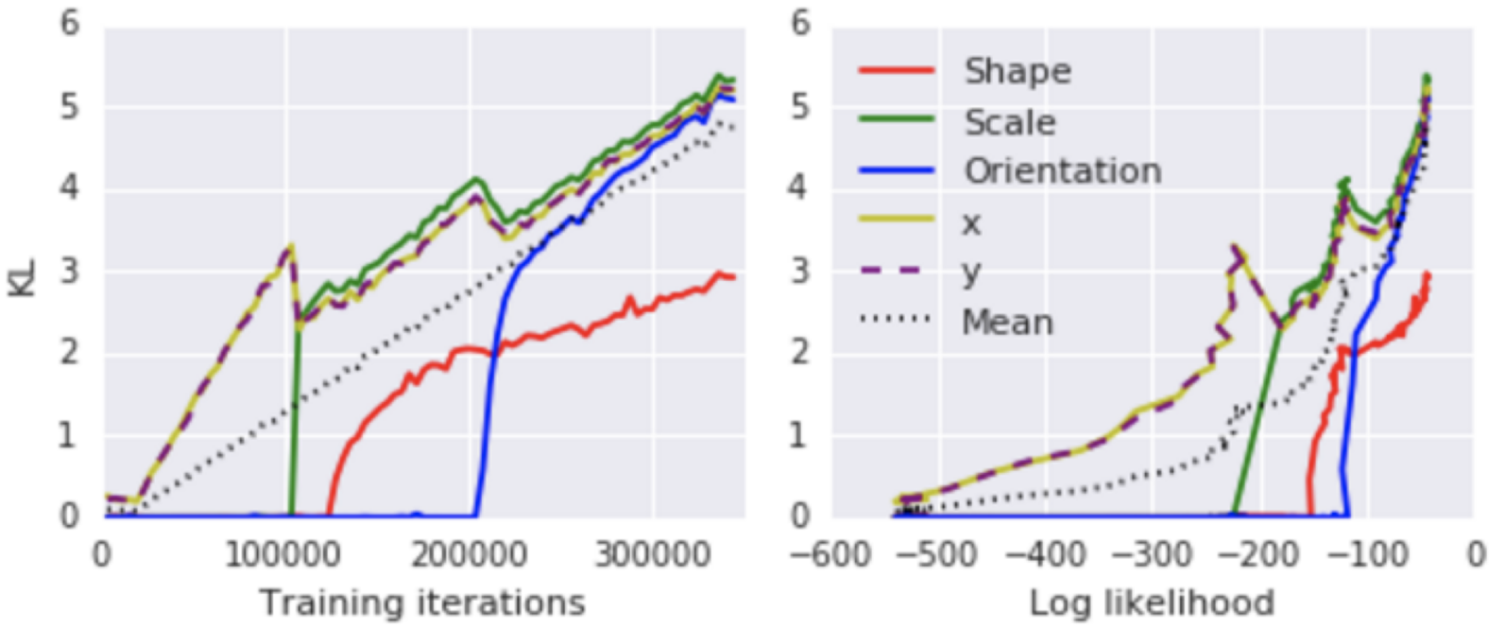


Burgess C. P. et al. Understanding disentangling in β-VAE, 2018

β-VAE

Controlled encoding capacity

$$\mathcal{L}(q, \boldsymbol{\theta}, \beta) = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \log p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) - |KL(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) - C|.$$



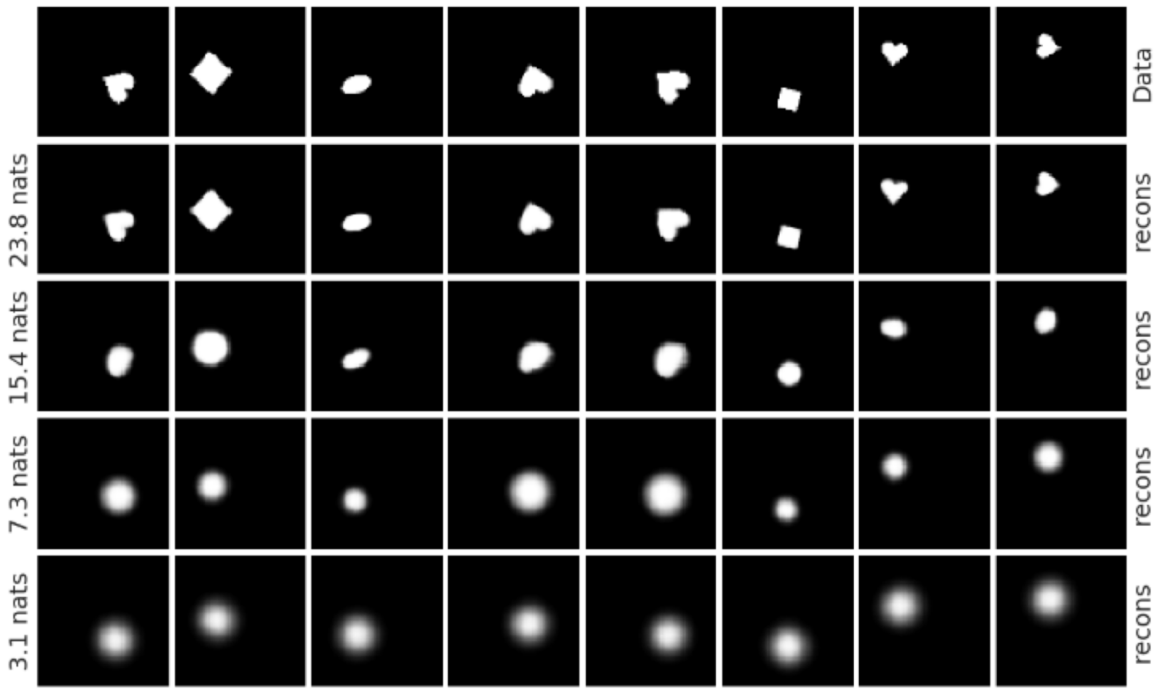
The early capacity is allocated to positional latents only, followed by a scale latent, then shape and orientation latents.

Burgess C. P. et al. Understanding disentangling in β-VAE, 2018

β-VAE

Controlled encoding capacity

$$\mathcal{L}(q, \boldsymbol{\theta}, \beta) = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \log p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) - |KL(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) - C|.$$



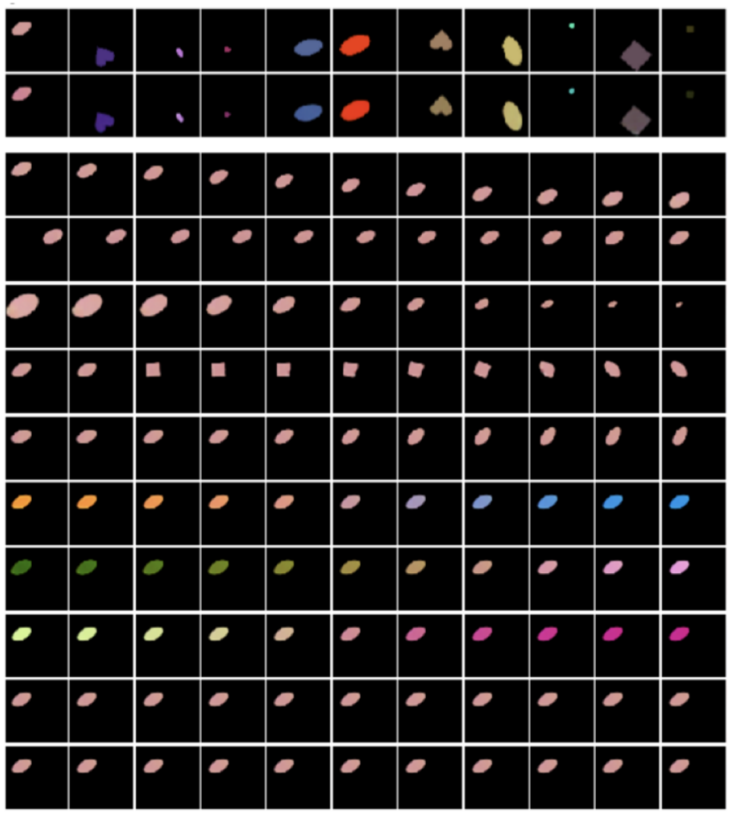
As the information capacity increases the different latents associated with their data generative factors become informative.

Burgess C. P. et al. Understanding disentangling in β-VAE, 2018

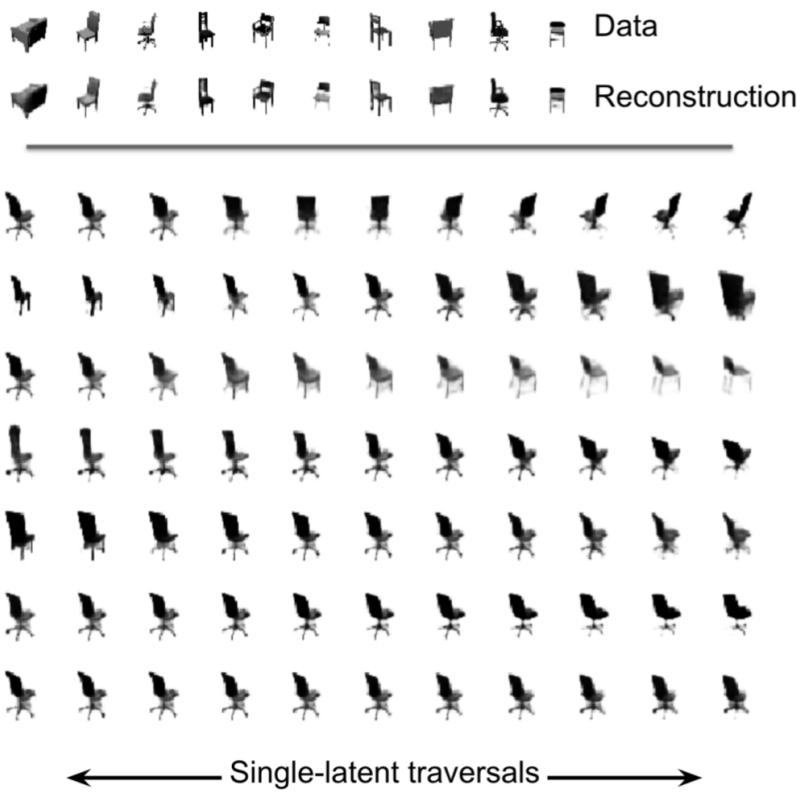
β-VAE

Single latent traversals, ordered by their average KL divergence with the prior

(a) Coloured dSprites



(b) 3D Chairs



Burgess C. P. et al. Understanding disentangling in β-VAE, 2018

β-VAE

ELBO

$$\mathcal{L}(q, \theta) = \frac{1}{n} \sum_{i=1}^{n} \left[\mathbb{E}_{q(\mathbf{z}_i|\mathbf{x}_i)} \log p(\mathbf{x}_i|\mathbf{z}_i, \theta) - \beta \cdot KL(q(\mathbf{z}_i|\mathbf{x}_i)||p(\mathbf{z}_i)) \right].$$

ELBO surgery

$$\mathcal{L}(q, \theta) = \underbrace{\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{q(\mathbf{z}_i|\mathbf{x}_i)} \log p(\mathbf{x}_i|\mathbf{z}_i, \theta)}_{\text{Reconstruction loss}} - \beta \cdot \underbrace{\mathbb{I}_{q(i,\mathbf{z})}[i, \mathbf{z}]}_{\text{Mutual info}} - \beta \cdot \underbrace{KL(q(\mathbf{z})||p(\mathbf{z}))}_{\text{Marginal KL}}$$

Minimization of MI

- It is not necessary and not desirable for disentanglement.
- It hurts reconstruction.

Burgess C. P. et al. Understanding disentangling in β-VAE, 2018

DIP-VAE

Disentangled aggregated variational posterior

$$q(\mathbf{z}) = \mathbb{E}_{\pi(\mathbf{x})} q(\mathbf{z}|\mathbf{x}) = \int q(\mathbf{z}|\mathbf{x})\pi(\mathbf{x})d\mathbf{x} = \prod_{j=1}^{d} q(z_j)$$

Variational inference with disentangled prior encourages inferring factors that are close to being disentangled:

$$KL(q(\mathbf{z})||\mathbb{E}_{\pi(\mathbf{x})} p(\mathbf{z}|\mathbf{x})) \leq \mathbb{E}_{\pi(\mathbf{x})} KL(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z}|\mathbf{x}))$$

DIP-VAE Objective

$$\mathcal{L}(q, \boldsymbol{\theta}) = \underbrace{\mathbb{E}_{\pi(\mathbf{x})}\left[\mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \log p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) - KL(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))\right]}_{\text{ELBO}} - \lambda \cdot KL(q(\mathbf{z})||p(\mathbf{z})) =$$

$$= \underbrace{\mathbb{E}_{\pi(\mathbf{x})}\left[\mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \log p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta})\right]}_{\text{Reconstruction loss}} - \underbrace{\mathbb{I}_{q(i,\mathbf{z})}[i, \mathbf{z}]}_{\text{Mutual info}} - (1 + \lambda) \cdot \underbrace{KL(q(\mathbf{z})||p(\mathbf{z}))}_{\text{Marginal KL}}$$

Kumar A., Sattigeri P., Balakrishnan A. Variational Inference of Disentangled Latent Concepts from Unlabeled Observations, 2017

DIP-VAE

DIP-VAE Objective

$$\mathcal{L}(q, \boldsymbol{\theta}) = \underbrace{\mathbb{E}_{\pi(\mathbf{x})}\left[\mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \log p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) - KL(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))\right]}_{\text{ELBO}} - \lambda \cdot KL(q(\mathbf{z})||p(\mathbf{z}))$$

- $KL(q(\mathbf{z})||p(\mathbf{z}))$ is intractable.
- Let match the moments of $q(\mathbf{z})$ and $p(\mathbf{z})$.

$$\text{cov}_{q(\mathbf{z})}(\mathbf{z}) = \mathbb{E}_{\pi(\mathbf{x})}\text{cov}_{q(\mathbf{z}|\mathbf{x})}(\mathbf{z}) + \text{cov}_{\pi(\mathbf{x})}\left(\mathbb{E}_{q(\mathbf{z}|\mathbf{x})}\mathbf{z}\right).$$

For most common case $q(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}(\mathbf{x}), \boldsymbol{\Sigma}(\mathbf{x}))$:

$$\text{cov}_{q(\mathbf{z})}(\mathbf{z}) = \mathbb{E}_{\pi(\mathbf{x})}\boldsymbol{\Sigma}(\mathbf{x}) + \text{cov}_{\pi(\mathbf{x})}\boldsymbol{\mu}(\mathbf{x})$$

DIP-VAE regularizes $\text{cov}_{q(\mathbf{z})}(\mathbf{z})$ to be close to the identity matrix.

Kumar A., Sattigeri P., Balakrishnan A. Variational Inference of Disentangled Latent Concepts from Unlabeled Observations, 2017

DIP-VAE

DIP-VAE Objective

$$\mathcal{L}(q, \theta) = \underbrace{\mathbb{E}_{\pi(\mathbf{x})} \left[\mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \log p(\mathbf{x}|\mathbf{z}, \theta) - KL(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) \right]}_{\text{ELBO}} - \lambda \cdot KL(q(\mathbf{z})||p(\mathbf{z}))$$

$$\text{cov}_{q(\mathbf{z})}(\mathbf{z}) = \mathbb{E}_{\pi(\mathbf{x})} \boldsymbol{\Sigma}(\mathbf{x}) + \text{cov}_{\pi(\mathbf{x})} \boldsymbol{\mu}(\mathbf{x})$$

DIP-VAE-I

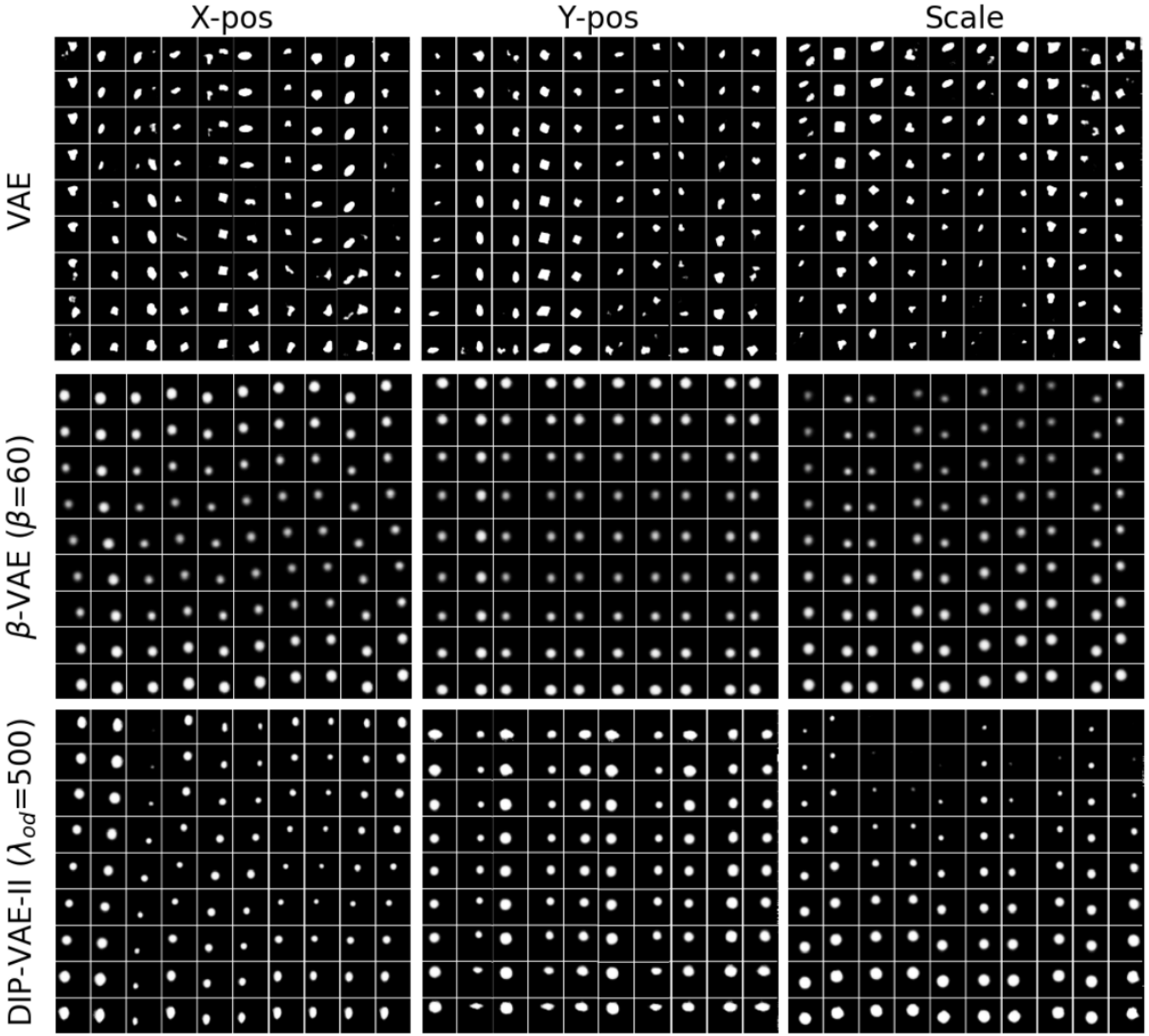
$$\max_{\theta, \phi} \text{ELBO}(\theta, \phi) - \lambda_1 \sum_{i \neq j} \left[\text{cov}_{\pi(\mathbf{x})} \boldsymbol{\mu}(\mathbf{x}) \right]_{ij}^2 - \lambda_2 \sum_i \left(\left[\text{cov}_{\pi(\mathbf{x})} \boldsymbol{\mu}(\mathbf{x}) \right]_{ii} - 1 \right)^2$$

DIP-VAE-II

$$\max_{\theta, \phi} \text{ELBO}(\theta, \phi) - \lambda_1 \sum_{i \neq j} \left[\text{cov}_{q(\mathbf{z})}(\mathbf{z}) \right]_{ij}^2 - \lambda_2 \sum_i \left(\left[\text{cov}_{q(\mathbf{z})}(\mathbf{z}) \right]_{ii} - 1 \right)^2$$

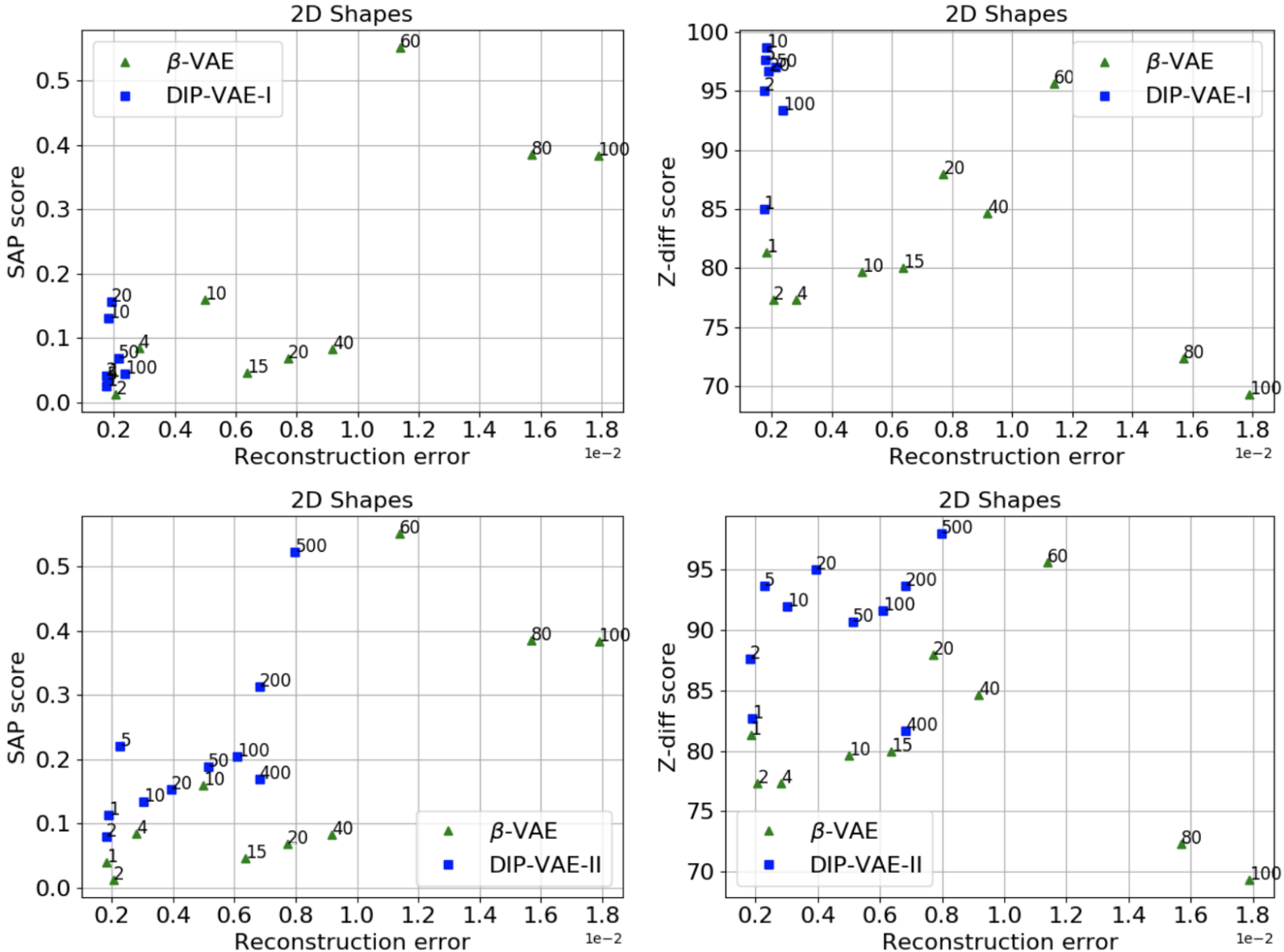
Kumar A., Sattigeri P., Balakrishnan A. Variational Inference of Disentangled Latent Concepts from Unlabeled Observations, 2017

DIP-VAE



https://arxiv.org/abs/1711.00848

DIP-VAE



Kumar A., Sattigeri P., Balakrishnan A. Variational Inference of Disentangled Latent Concepts from Unlabeled Observations, 2017

FactorVAE

Disentangled aggregated variational posterior

$$q(\mathbf{z}) = \mathbb{E}_{\pi(\mathbf{x})} q(\mathbf{z}|\mathbf{x}) = \int q(\mathbf{z}|\mathbf{x})\pi(\mathbf{x})d\mathbf{x} = \prod_{j=1}^{d} q(z_j)$$

Total correlation regularizer

$$\min KL(q(\mathbf{z})||\prod_{j=1}^{d} q(z_j))$$

FactorVAE objective

$$\min_{\boldsymbol{\theta},\boldsymbol{\phi}} \text{ELBO}(\boldsymbol{\theta},\boldsymbol{\phi}) - \gamma \cdot KL(q(\mathbf{z})||\prod_{j=1}^{d} q(z_j))$$

- The last term is intractable.
- FactorVAE uses density ratio trick for estimation.

Kim H., Mnih A. Disentangling by Factorising, 2018

FactorVAE

Consider two distributions $q_1(\mathbf{x})$, $q_2(\mathbf{x})$ and probilistic model

$$p(\mathbf{x}|y) = \begin{cases} q_1(\mathbf{x}), & \text{if } y = 1, \\ q_2(\mathbf{x}), & \text{if } y = 0, \end{cases} \qquad y \sim \text{Bern}(0.5).$$

Density ratio trick

$$\frac{q_1(\mathbf{x})}{q_2(\mathbf{x})} = \frac{p(\mathbf{x}|y=1)}{p(\mathbf{x}|y=0)} = \frac{p(y=1|\mathbf{x})p(\mathbf{x})}{p(y=1)} \Big/ \frac{p(y=0|\mathbf{x})p(\mathbf{x})}{p(y=0)} =$$

$$= \frac{p(y=1|\mathbf{x})}{p(y=0|\mathbf{x})} = \frac{p(y=1|\mathbf{x})}{1-p(y=1|\mathbf{x})} = \frac{D(\mathbf{x})}{1-D(\mathbf{x})}$$

Here $D(\mathbf{x})$ could be treated as a discriminator a model the output of which is a probability that $\mathbf{x}$ is a sample from $q_1(\mathbf{x})$ rather than from $q_2(\mathbf{x})$.

Kim H., Mnih A. Disentangling by Factorising, 2018

FactorVAE

FactorVAE objective

$$\min_{\boldsymbol{\theta},\boldsymbol{\phi}} \text{ELBO}(\boldsymbol{\theta},\boldsymbol{\phi}) - \gamma \cdot KL(q(\mathbf{z})||\prod_{j=1}^{d} q(z_j))$$

Total correlation regularizer

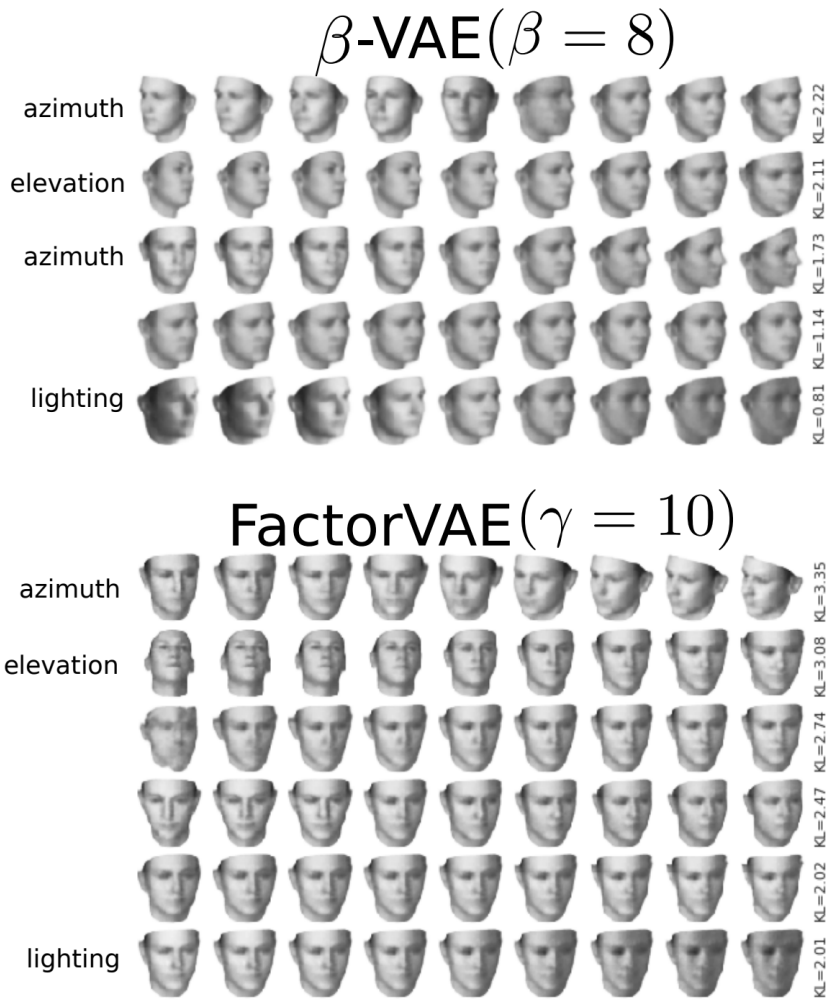
$$KL(q(\mathbf{z})||\prod_{j=1}^{d} q(z_j)) = KL(q(\mathbf{z})||\bar{q}(\mathbf{z})) =$$

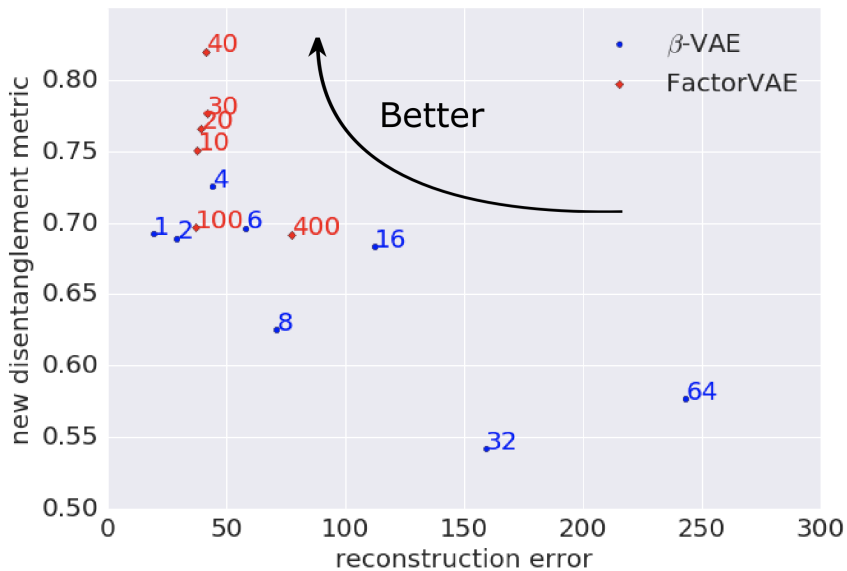
$$= \mathbb{E}_{q(\mathbf{z})} \log \frac{q(\mathbf{z})}{\bar{q}(\mathbf{z})} \approx \mathbb{E}_{q(\mathbf{z})} \log \frac{D(\mathbf{z})}{1 - D(\mathbf{z})}$$

VAE and GAN are trained simultaneously.

Kim H., Mnih A. Disentangling by Factorising, 2018

FactorVAE





Kim H., Mnih A. Disentangling by Factorising, 2018

Challenging Disentanglement Assumptions

Whether unsupervised disentanglement learning is even possible for arbitrary generative models?

Theorem

For $d > 1$, let $\mathbf{z} \sim P$ denote any distribution which admits a density $p(\mathbf{z}) = \prod_{i=1}^{d} p(z_i)$. Then, there exists an infinite family of bijective functions $f : \text{supp}(\mathbf{z}) \to \text{supp}(\mathbf{z})$ such that

- $\frac{\partial f_i(\mathbf{u})}{\partial u_j} \neq 0$ almost everywhere for all i and j (i.e., $\mathbf{z}$ and $f(\mathbf{z})$ are completely entangled);
- and $P(\mathbf{z} \leq \mathbf{u}) = P(f(\mathbf{z}) \leq \mathbf{u})$ for all $\mathbf{u} \in \text{supp}(\mathbf{z})$ (i.e., they have the same marginal distribution).

Theorem claims that unsupervised disentanglement learning is impossible for arbitrary generative models with a factorized prior.

Locatello F. et al. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations, 2018

Challenging Disentanglement Assumptions

Assume we have $p(\mathbf{z})$ and some $p(\mathbf{x}|\mathbf{z})$ defining a generative model. Consider any unsupervised disentanglement method and assume that it finds a representation that is perfectly disentangled with respect to $\mathbf{z}$ in the generative model.

- Theorem claims that $\exists\ \hat{\mathbf{z}} = f(\mathbf{z})$ where $\hat{\mathbf{z}}$ is completely entangled with respect to $\mathbf{z}$.
- Since the (unsupervised) disentanglement method only has access to observations $\mathbf{x}$, it hence cannot distinguish between the two equivalent generative models and thus has to be entangled to at least one of them

$$p(\mathbf{x}) = \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z})d\mathbf{z} = \int p(\mathbf{x}|\hat{\mathbf{z}})p(\hat{\mathbf{z}})d\hat{\mathbf{z}}.$$

Locatello F. et al. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations, 2018

Challenging Disentanglement Assumptions

Proof (1)

1. Consider the function $g : \text{supp}(\mathbf{z}) \to [0,1]^d$:

$$g_i(\mathbf{v}) = P(z_i \leq v_i), \quad i = 1, \ldots, d.$$

   - g is bijective (since $p(\mathbf{z}) = \prod_{i=1} dp(z_i)$).
   - $\frac{\partial g_i(\mathbf{u})}{\partial u_i} \neq 0$, for all i and $\frac{\partial g_i(\mathbf{u})}{\partial u_j} = 0$ for all $i \neq j$.
   - $g(\mathbf{z})$ is an independent d-dimensional uniform distribution.

2. Consider $h : (0,1]^d \to \mathbb{R}^d$

$$h_i(\mathbf{v}) = \psi^{-1}(v_i), \quad i = 1, \ldots, d.$$

   Here ψ denotes the CDF of a standard normal distribution.

   - h is bijective.
   - $\frac{\partial g_i(\mathbf{u})}{\partial u_i} \neq 0$, for all i and $\frac{\partial g_i(\mathbf{u})}{\partial u_j} = 0$ for all $i \neq j$.
   - $h(g(\mathbf{z}))$ is a d-dimensional standard normal distribution.

Locatello F. et al. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations, 2018

Challenging Disentanglement Assumptions

Proof (2)

Let $\mathbf{A} \in \mathbb{R}^{d \times d}$ be an arbitrary orthogonal matrix with $A_{ij} \neq 0$ for all i, j. The family of such matrices is infinite.

- $\mathbf{A}$ is orthogonal, it is invertible and thus defines a bijective linear operator.
- $\mathbf{A}h(g(\mathbf{z})) \in \mathbb{R}^d$ is hence an independent, multivariate standard normal distribution.
- $h^{-1}(\mathbf{A}h(g(\mathbf{z}))) \in \mathbb{R}^d$ is an independent d-dimensional uniform distribution.

Define $f : \text{supp}(\mathbf{z}) \to \text{supp}(\mathbf{z})$:

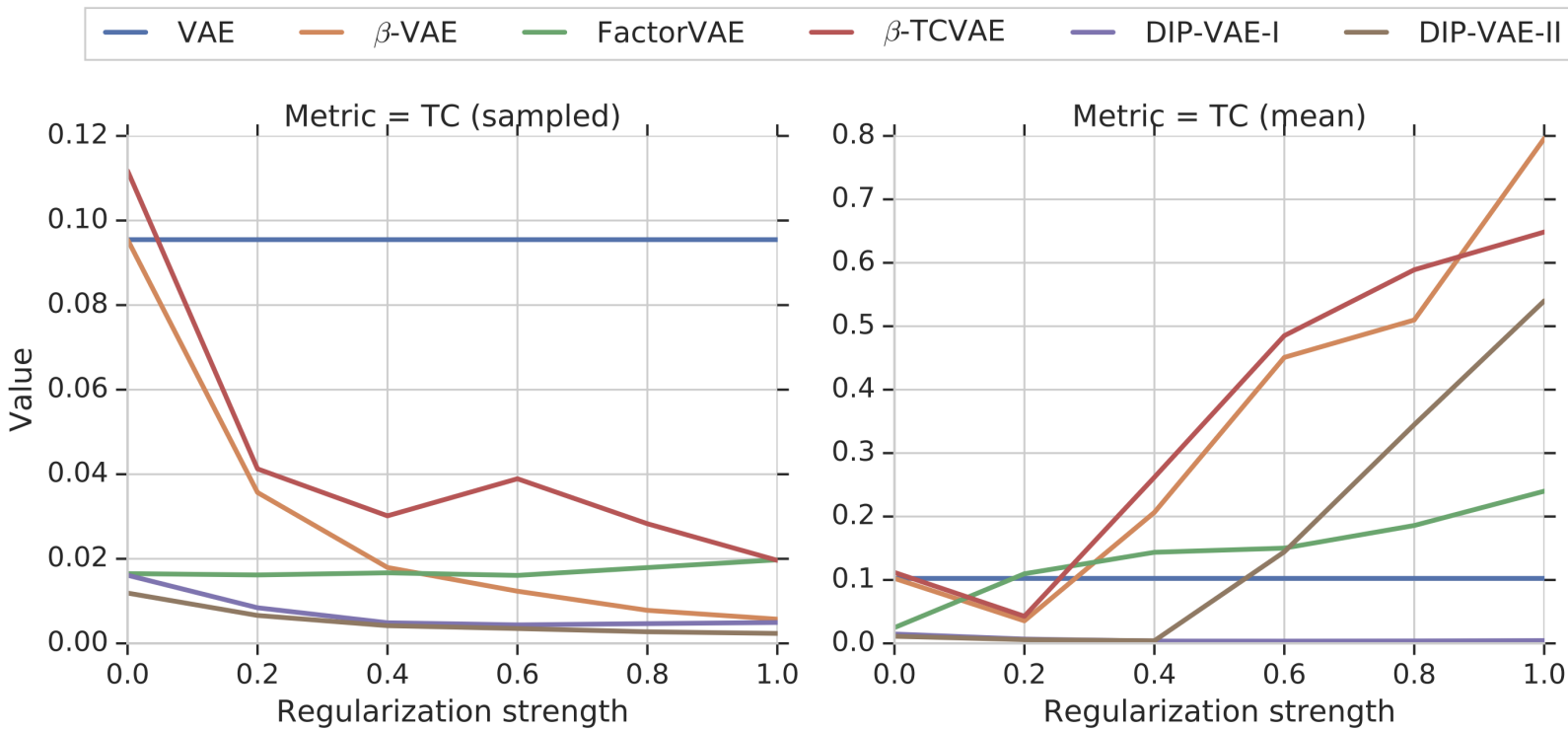
$$f(\mathbf{u}) = g^{-1}(h^{-1}(\mathbf{A}h(g(\mathbf{z})))).$$

By definition $f(\mathbf{z})$ has the same marginal distribution as $\mathbf{z}$:
$P(\mathbf{z} \leq \mathbf{u}) = P(f(\mathbf{z}) \leq \mathbf{u})$ and $\frac{\partial f_i(\mathbf{u})}{\partial u_j} \neq 0$.

Locatello F. et al. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations, 2018

Challenging Disentanglement Assumptions

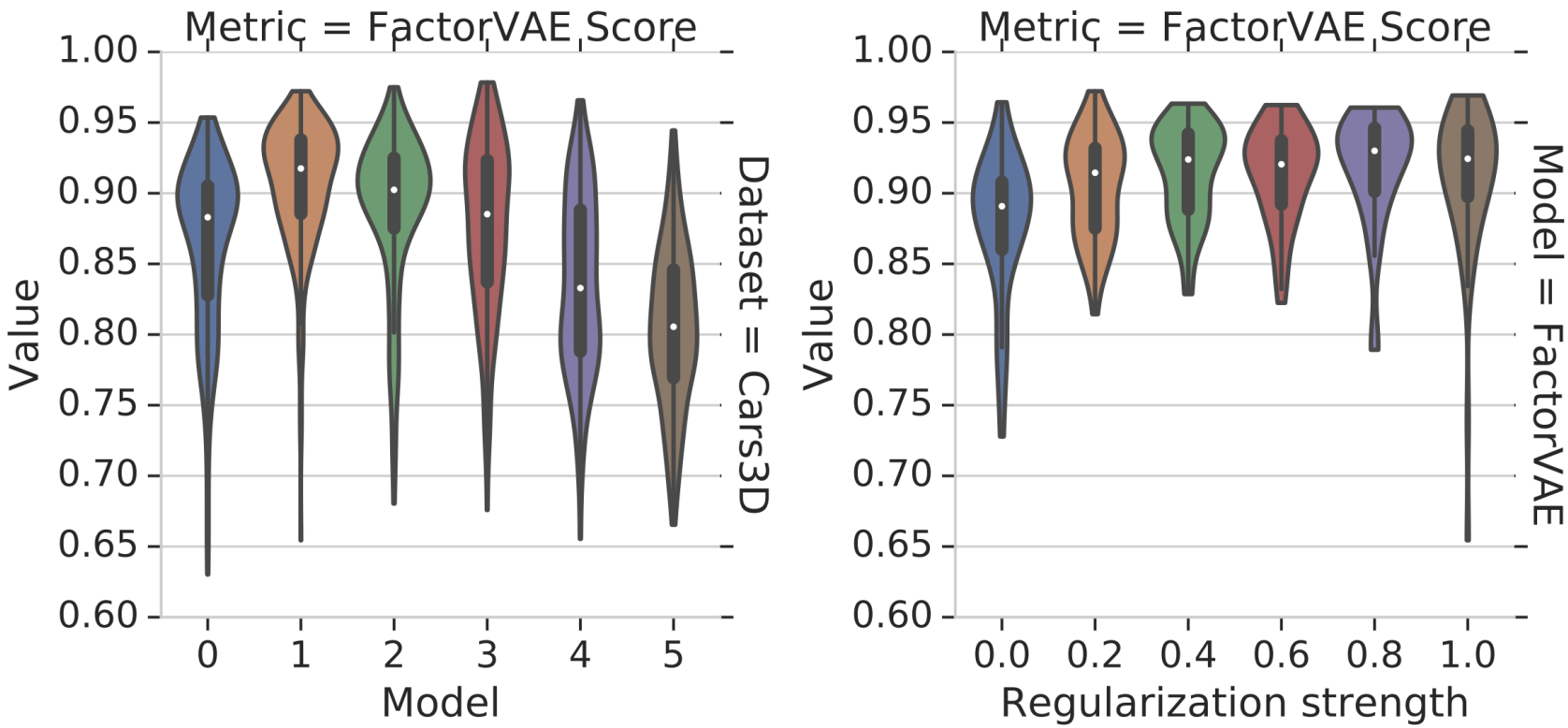
- **Training:** Factorizing **samples** from aggregated posterior $q(\mathbf{z}) = \prod_{i=1}^{d} q(z_i)$.
- **Inference:** Use a **mean** vector (usually mean of Gaussian encoder) as a representation.



Locatello F. et al. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations, 2018

Challenging Disentanglement Assumptions

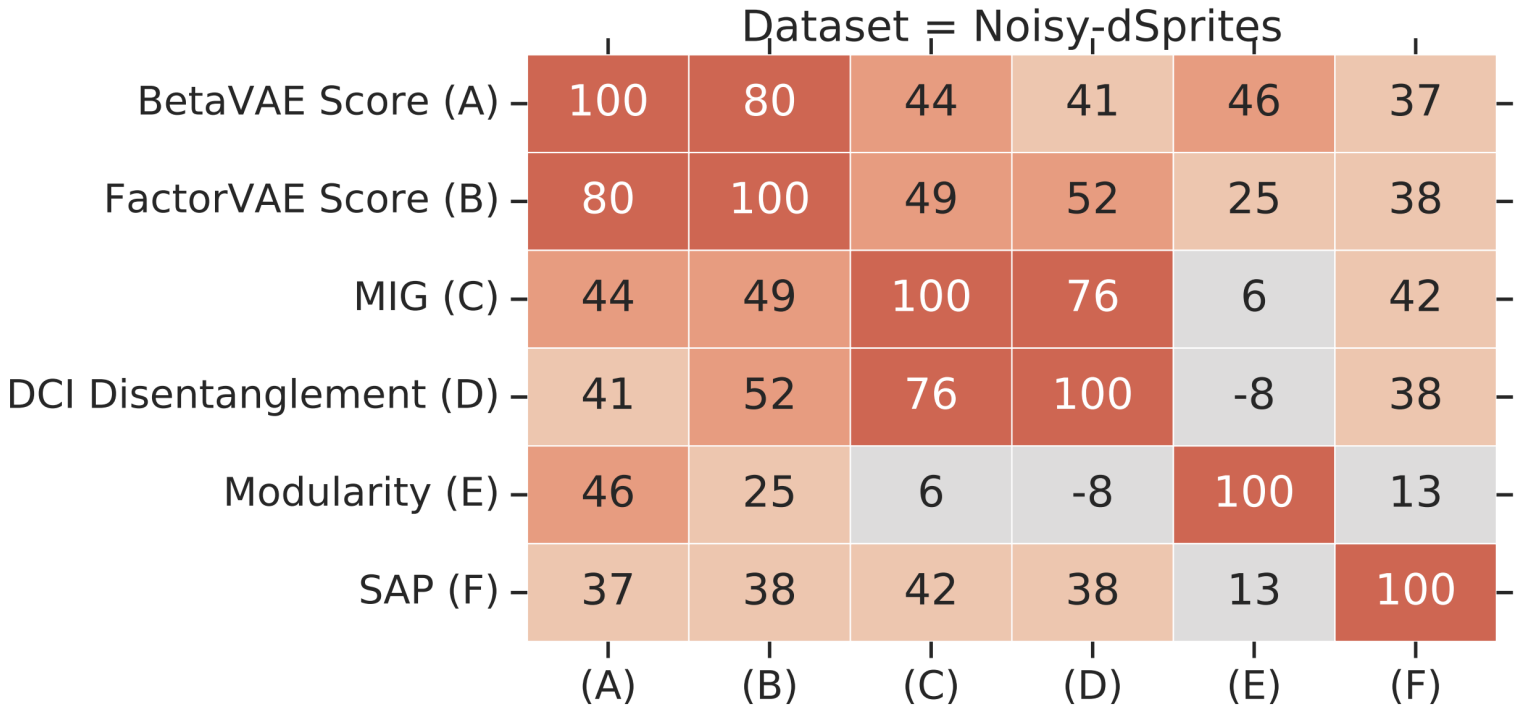
Importance of different models and hyperparameters for disentanglement



Locatello F. et al. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations, 2018

Challenging Disentanglement Assumptions

Agreement of different disentanglement metrics



Locatello F. et al. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations, 2018

Summary