Deep Generative Models
Lecture 11

Roman Isachenko

Moscow Institute of Physics and Technology

2020

Disentangled representations

Goal

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

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.

---

https://openreview.net/references/pdf?id=Sy2fzU9gl

β-VAE, 2017

Generative process

- $p(\mathbf{x}|\mathbf{v},\mathbf{w}) = \mathrm{Sim}(\mathbf{v},\mathbf{w})$ – true world simulator;
- $\mathbf{v}$ – conditionally independent factors: $p(\mathbf{v}|\mathbf{x}) = \prod_{k=1}^{K} p(v_k|\mathbf{x})$;
- $\mathbf{w}$ – conditionally dependent factors.

Goal

Develop 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 $\mathbf{v}$.

---

https://openreview.net/references/pdf?id=Sy2fzU9gl

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.

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.

https://arxiv.org/abs/1606.03657

InfoGAN

Objective

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

Variational Information Maximization

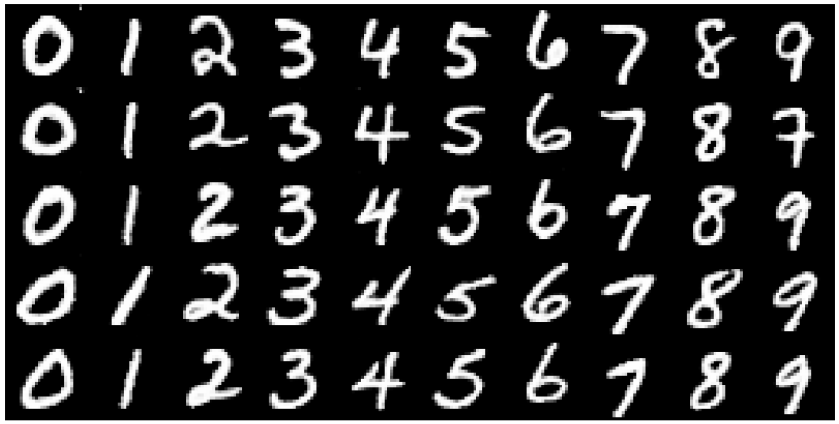
$$\begin{aligned} I(\mathbf{c}, G(\mathbf{z}, \mathbf{c})) &= H(\mathbf{c}) - H(\mathbf{c}|G(\mathbf{z}, \mathbf{c})) = \\ &= H(\mathbf{c}) + \mathbb{E}_{\mathbf{x}\sim G(\mathbf{z},\mathbf{c})}\left[\mathbb{E}_{\mathbf{c}\sim p(\mathbf{c}|\mathbf{x})} \log p(\mathbf{c}'|\mathbf{x})\right] = \\ &= H(\mathbf{c})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

---

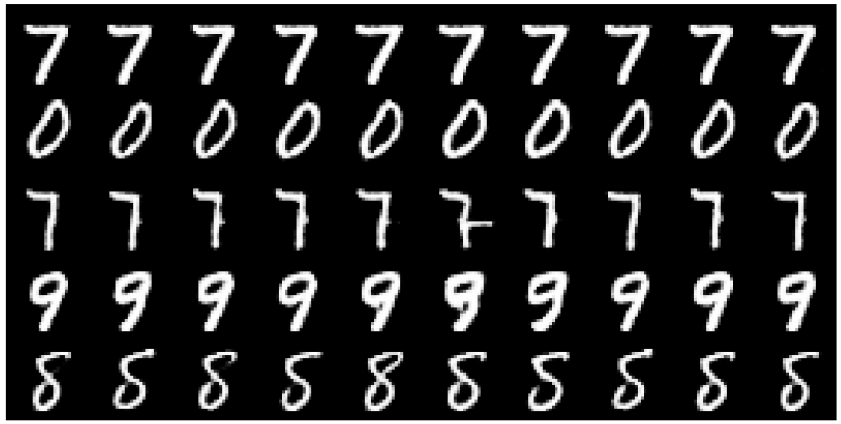
https://arxiv.org/abs/1606.03657

InfoGAN

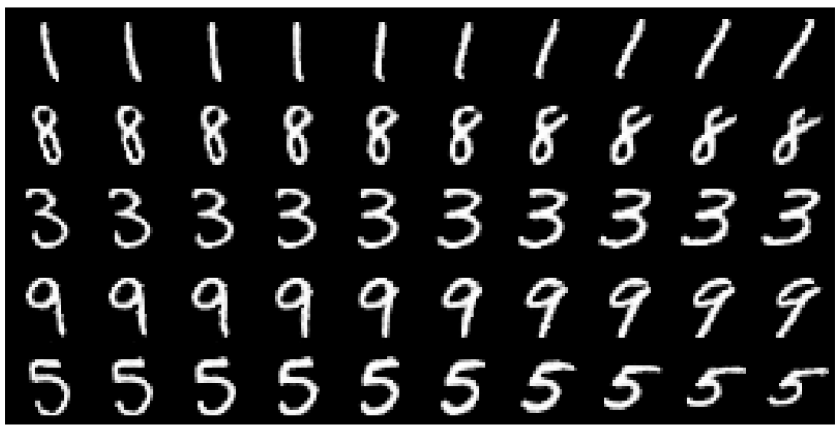
Latent codes on MNIST



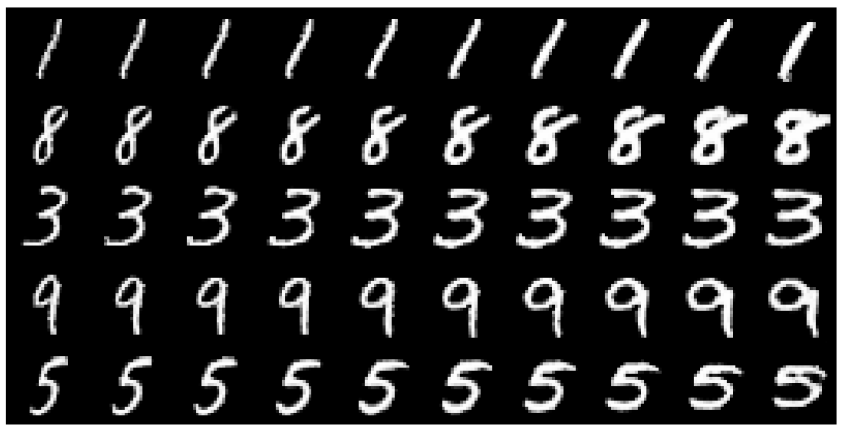
(a) Varying c_1 on InfoGAN (Digit type)



(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)

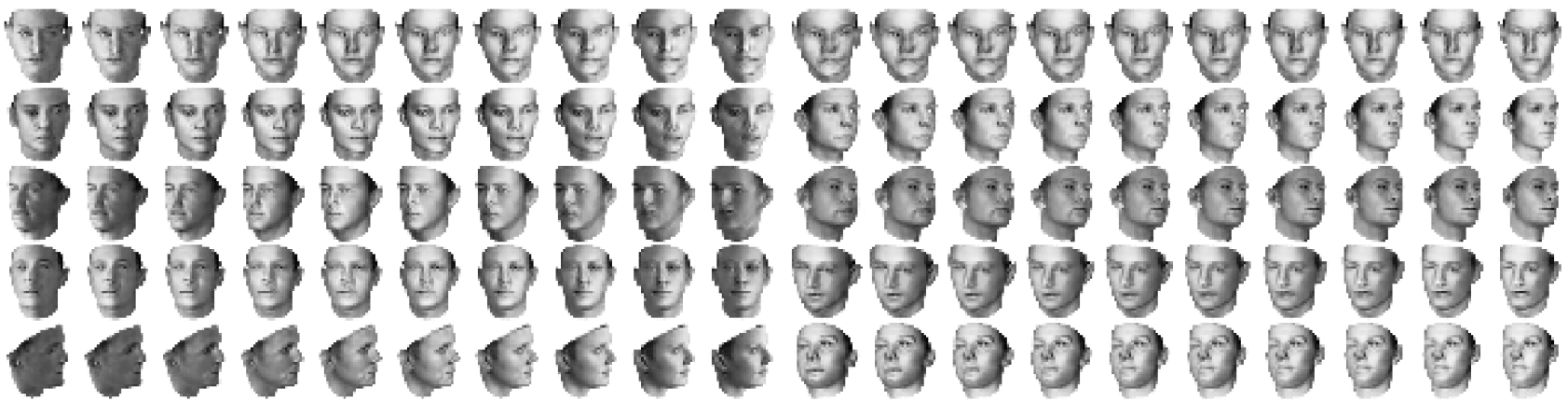
https://arxiv.org/abs/1606.03657

InfoGAN

Latent codes on 3D Faces

(a) Azimuth (pose)

(b) Elevation



(c) Lighting

(d) Wide or Narrow

https://arxiv.org/abs/1606.03657

β-VAE, 2017

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 stronger constraint on the latent bottleneck: $\beta > 1$.

Note: It could lead to poorer reconstructions due to the loss of high frequency details when passing through a constrained latent bottleneck.

---

https://openreview.net/references/pdf?id=Sy2fzU9gl

β-VAE, 2017

Disentangling metric

Accuracy of classifier $p(y|\mathbf{z}_{\text{diff}})$ with a low VC-dimension in order to ensure that it has no capacity to perform nonlinear disentangling itself.

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

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

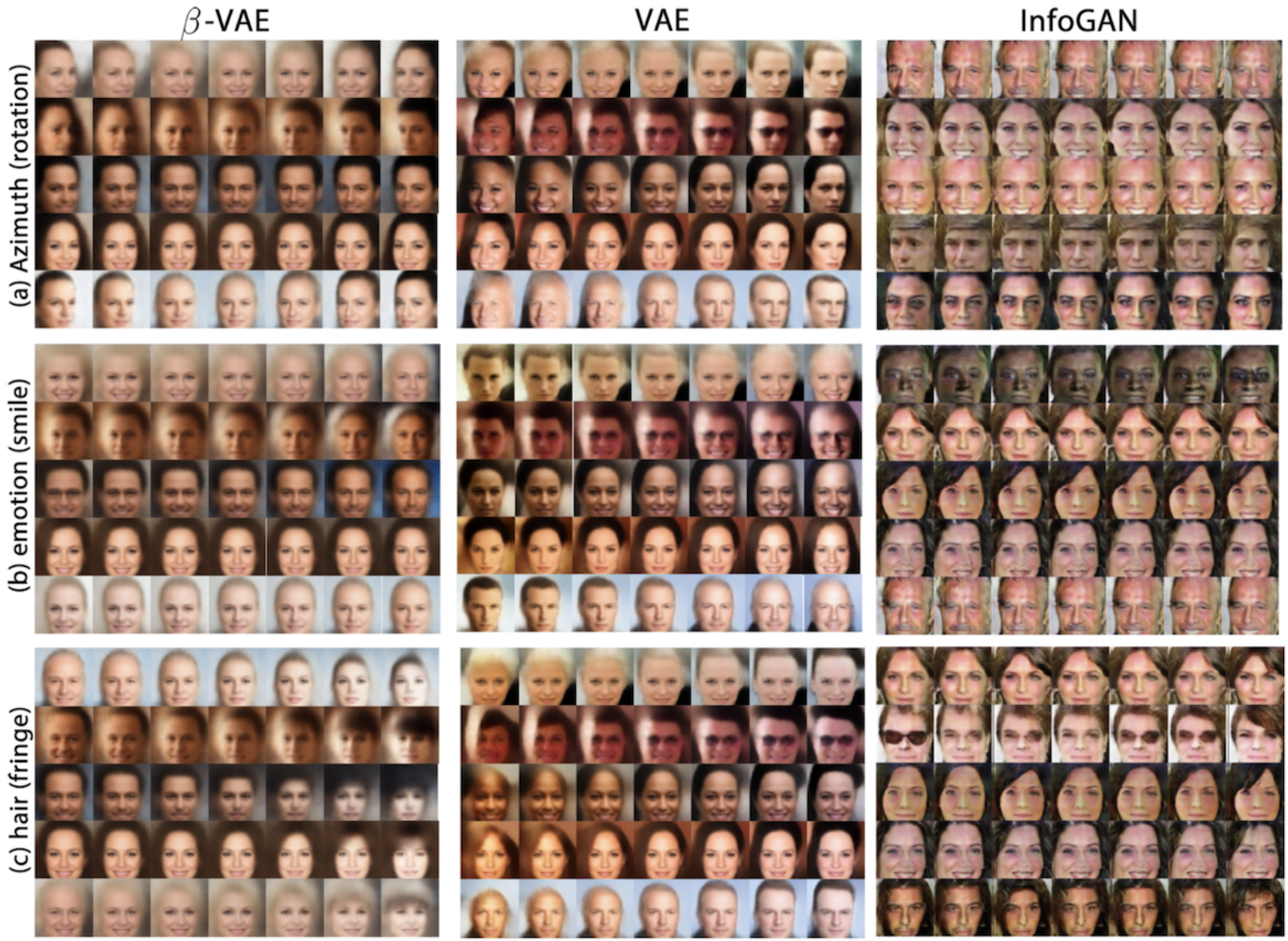
$$q(\mathbf{z}|\mathbf{x}) = \mathcal{N}\left(\mu(\mathbf{x})|\sigma^2(\mathbf{x})\right); \quad \mathbf{z}_{li} = \mu(\mathbf{x}_{li}); \quad \mathbf{z}_{lj} = \mu(\mathbf{x}_{lj}).$$

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

---

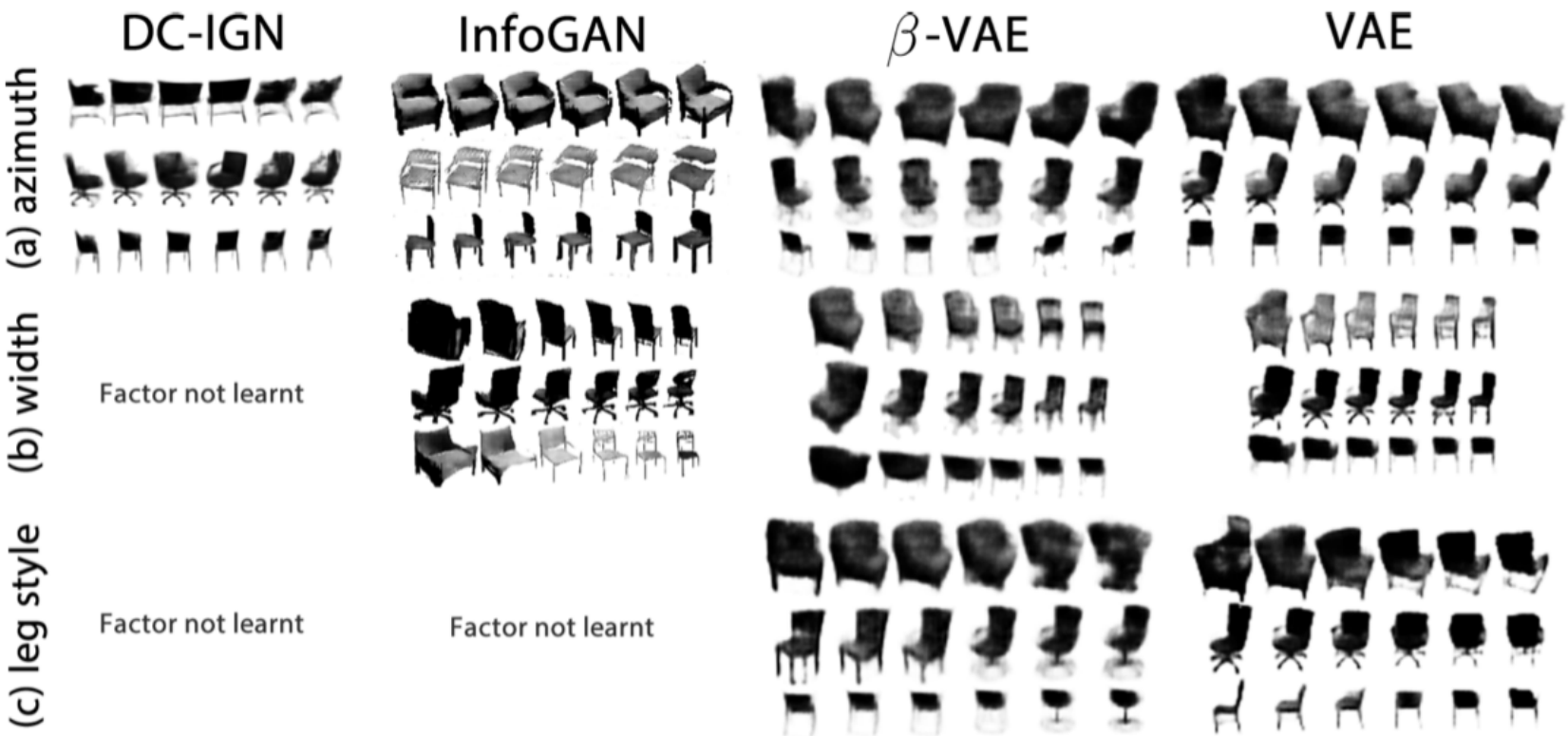
https://openreview.net/references/pdf?id=Sy2fzU9gl

β-VAE, 2017



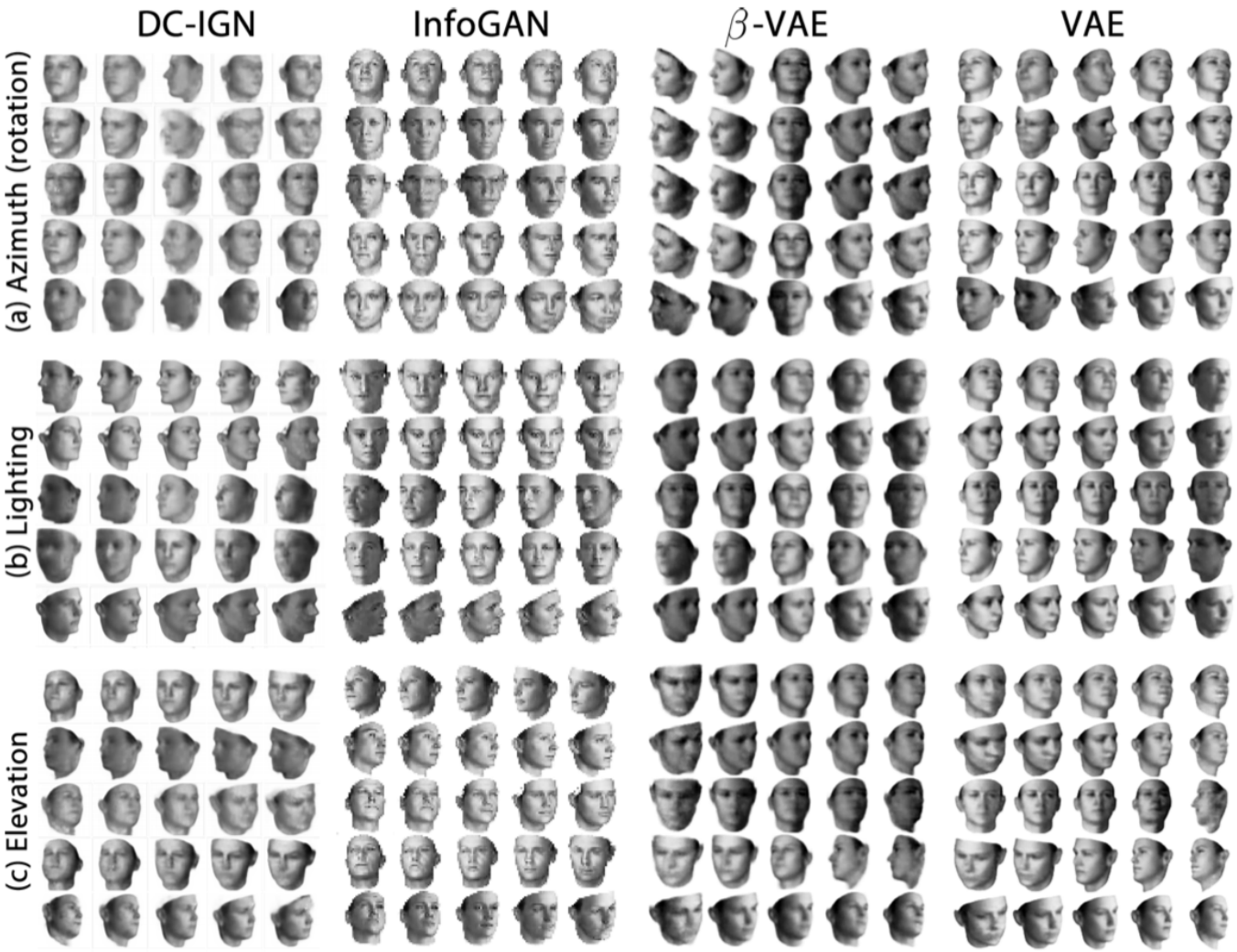
https://openreview.net/references/pdf?id=Sy2fzU9gl

β-VAE, 2017



https://openreview.net/references/pdf?id=Sy2fzU9gl

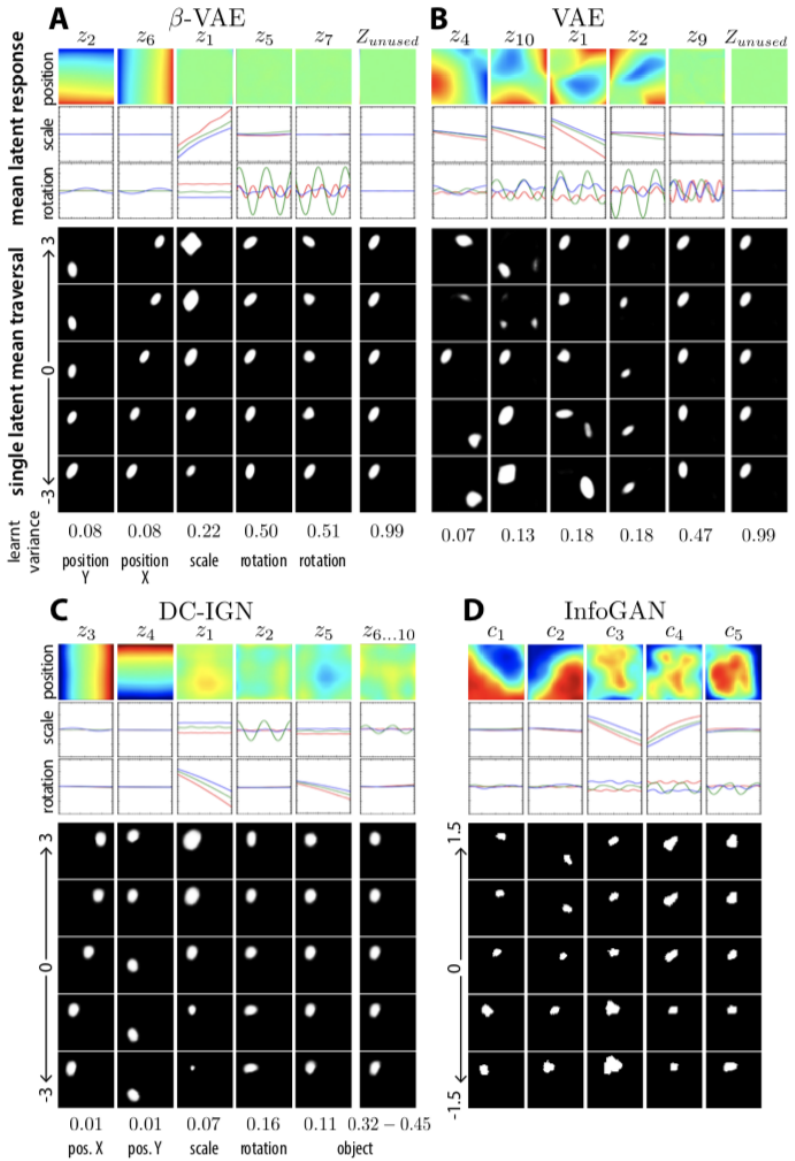
β-VAE, 2017



https://openreview.net/references/pdf?id=Sy2fzU9gl

β-VAE, 2017

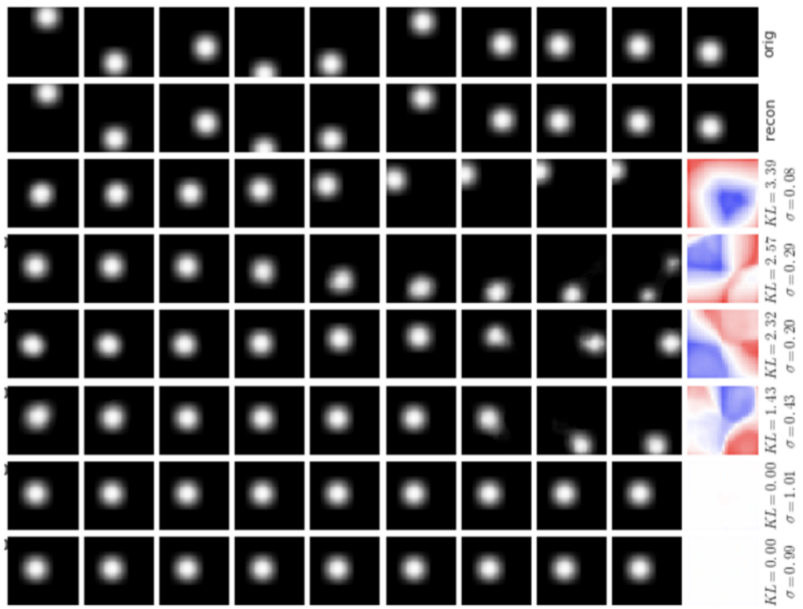
Model	Disentanglement metric score
Ground truth	*100%*
Raw pixels	$45.75 \pm 0.8\%$
PCA	$84.9 \pm 0.4\%$
ICA	$42.03 \pm 10.6\%$
DC-IGN	$\mathbf{99.3 \pm 0.1\%}$
InfoGAN	$73.5 \pm 0.9\%$
VAE untrained	$44.14 \pm 2.5\%$
VAE	$61.58 \pm 0.5\%$
β-VAE	$\mathbf{99.23 \pm 0.1\%}$



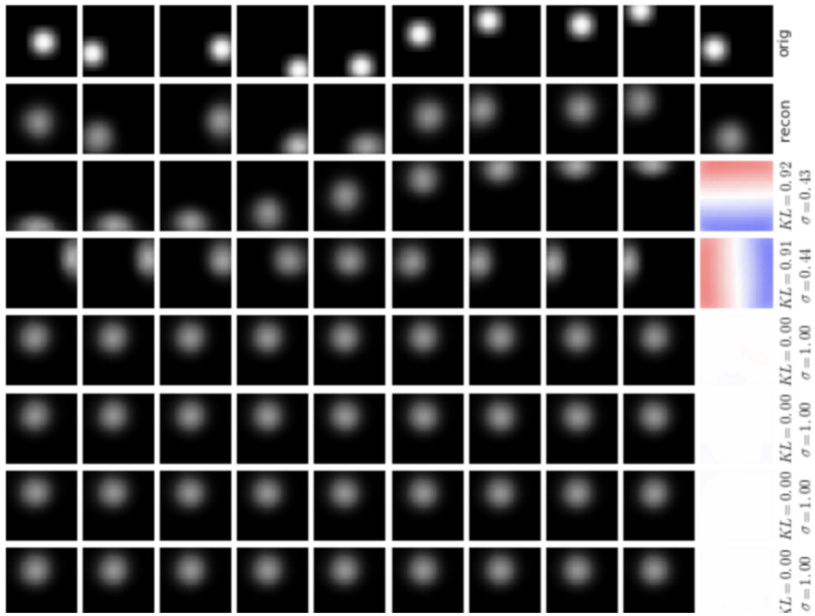
https://openreview.net/references/pdf?id=Sy2fzU9gl

β-VAE, 2018

β = 1



β = 150



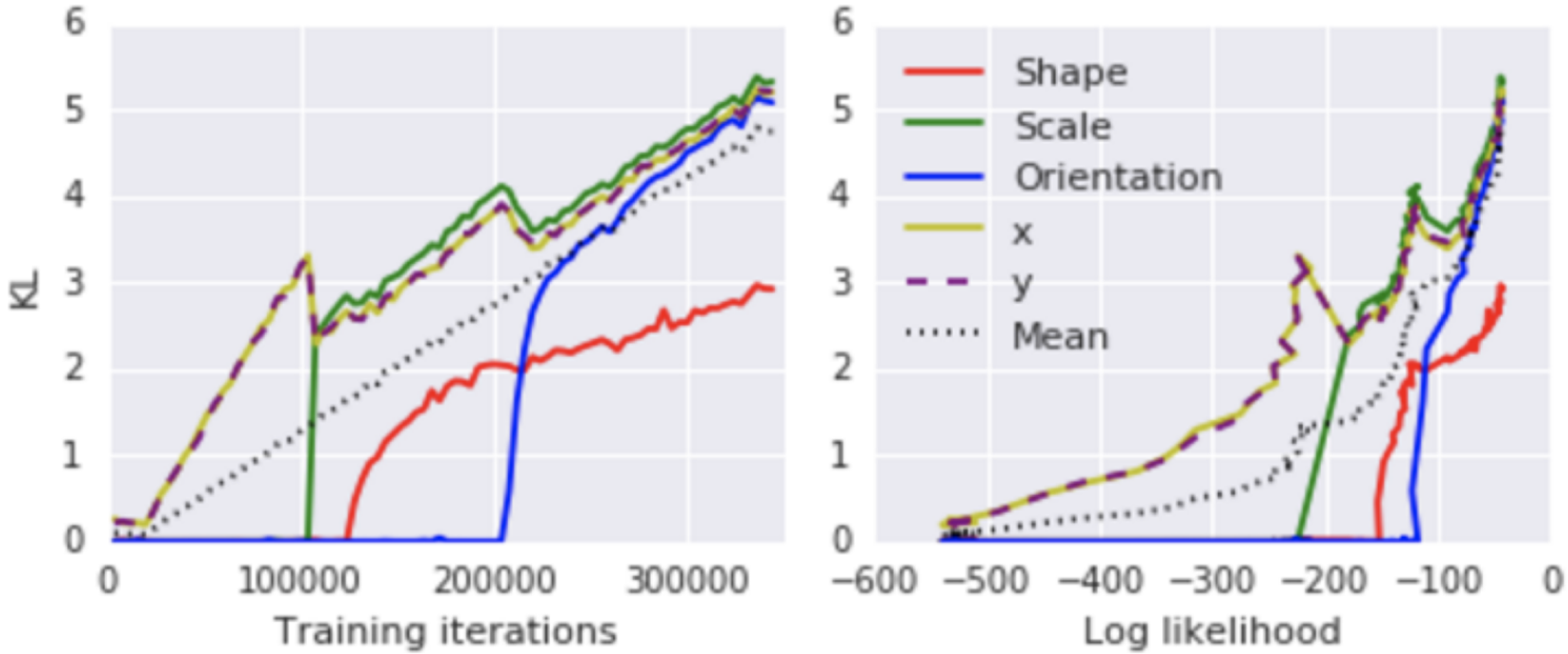
---

https://arxiv.org/pdf/1804.03599.pdf

β-VAE, 2018

Controlled encoding capacity

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

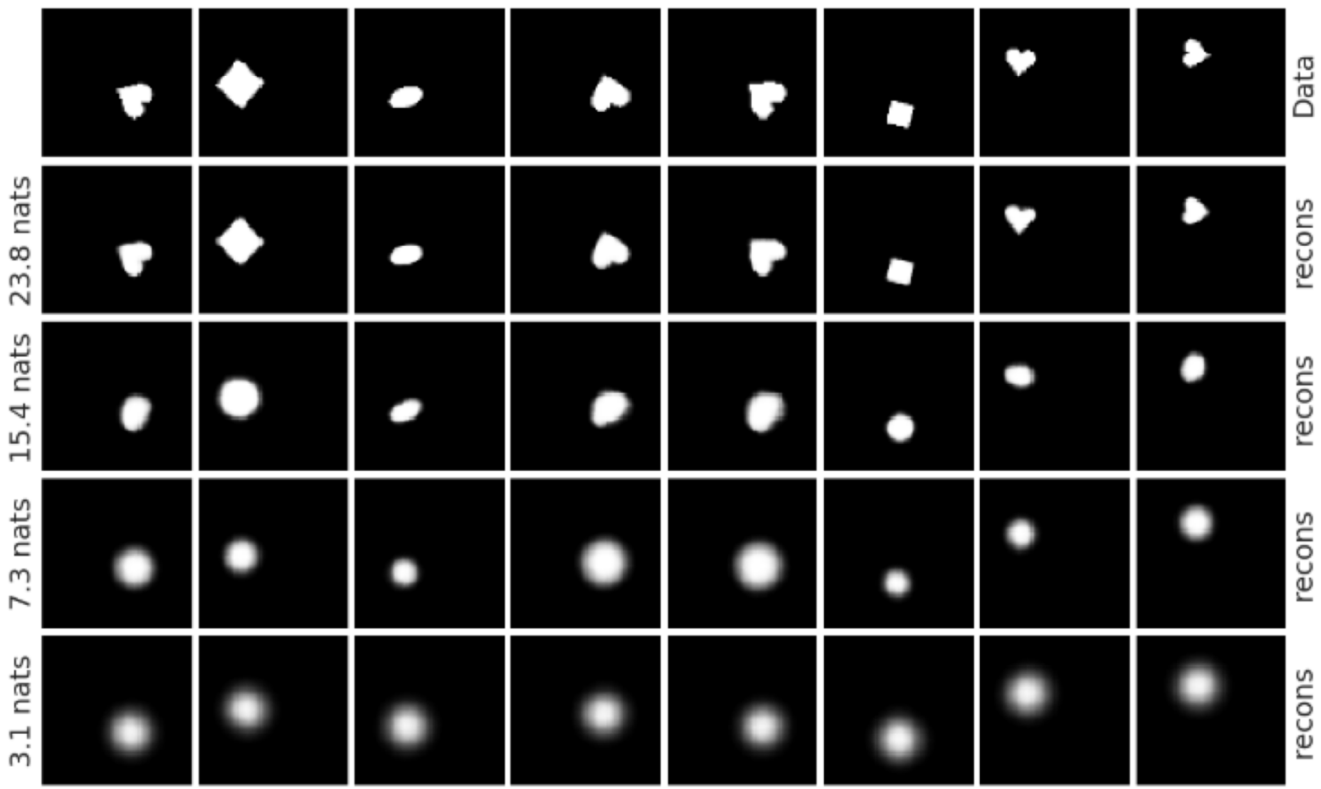


https://arxiv.org/pdf/1804.03599.pdf

β-VAE, 2018

Controlled encoding capacity

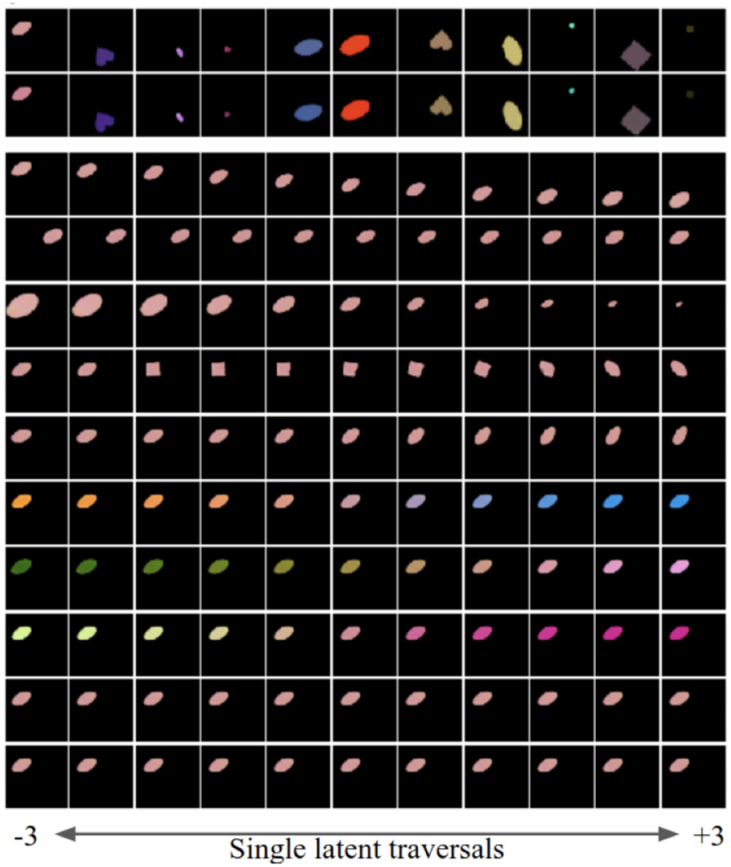
$$\mathcal{L}(q, \boldsymbol{\theta}, \beta) = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \log p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) - |KL(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) - C|.$$



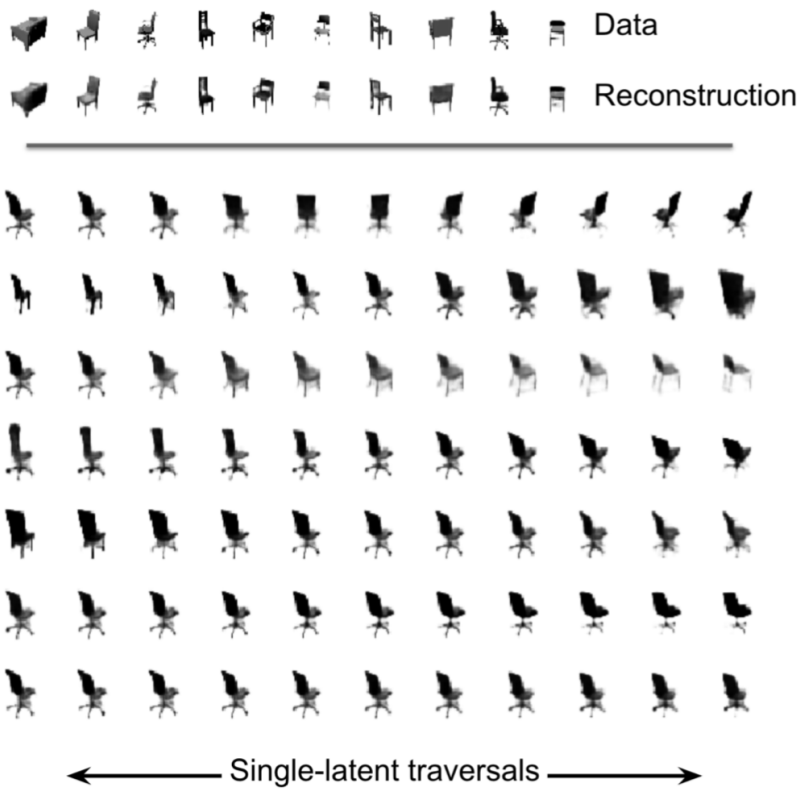
https://arxiv.org/pdf/1804.03599.pdf

β-VAE, 2018

(a) Coloured dSprites



(b) 3D Chairs



https://arxiv.org/pdf/1804.03599.pdf

DIP-VAE

---

https://arxiv.org/abs/1711.00848

FactorVAE

---

https://arxiv.org/abs/1802.05983

References

- **InfoGAN:** Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets
https://arxiv.org/abs/1606.03657
Summary: An information-theoretic extension to the GANs that disentangles representations in an unsupervised manner. InfoGAN maximizes the MI between a small subset of the latent variables and the observation. Lower bound for MI objective is derived that can be optimized efficiently.
- **beta-VAE:** Learning Basic Visual Concepts with a Constrained Variational Framework
https://openreview.net/references/pdf?id=Sy2fzU9gl
Summary: Modifications of VAE objective. The task is represented as constrained optimization. Increasing the weight of KL divergence term in ELBO allows to disentangle latent space factors and makes model more interpretable. The assessment of disentanglement is provided by constructing the classifier.
- Understanding disentangling in β-**VAE**
https://arxiv.org/pdf/1804.03599.pdf
Summary: Consider beta-VAE from the position of the rate-distortion theory (information bottleneck). Propose the modified ELBO with controlled latent capacity.
- **DIP-VAE:** Variational Inference of Disentangled Latent Concepts from Unlabeled Observations
https://arxiv.org/abs/1711.00848
Summary: Introduce a regularizer on the expectation of the approximate posterior over observed data that encourages the disentanglement. Penalize the mismatch between the aggregated posterior and a factorized prior. Comparison with beta-VAE.
- **FactorVAE:** Disentangling by Factorising
https://arxiv.org/abs/1802.05983
Summary: Penalizes the total correlation with density-ratio estimation. Comparison with beta-VAE. Does not degrate the reconstructions.