Deep Generative Models

Roman Isachenko

Moscow Institute of Physics and Technology

2019

VAE limitations

- Poor variational posterior distribution (encoder)

$$q(\mathbf{z}|\mathbf{x}, \phi) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_{\phi}(\mathbf{x}), \boldsymbol{\sigma}^2_{\phi}(\mathbf{x})).$$

- Poor prior distribution

$$p(\mathbf{z}) = \mathcal{N}(0, \mathbf{I}).$$

- Poor probabilistic model (decoder)

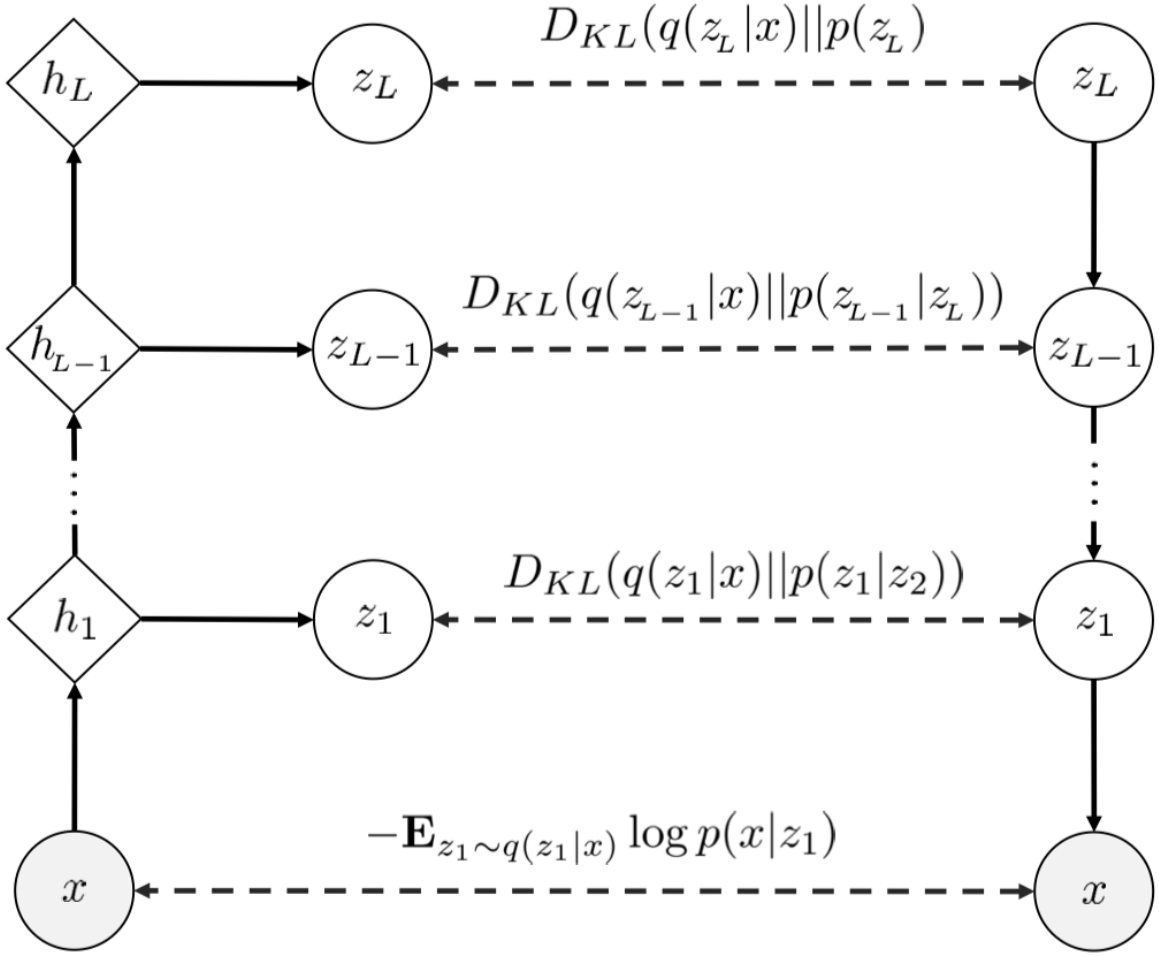
$$p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}), \boldsymbol{\sigma}^2_{\boldsymbol{\theta}}(\mathbf{z})).$$

- Loose lower bound

$$p(\mathbf{x}|\boldsymbol{\theta}) - \mathcal{L}(q, \boldsymbol{\theta}) = (?).$$

PixelVAE, 2016

Hierarchical VAE



https://arxiv.org/pdf/1611.05013.pdf

PixelVAE, 2016

Hierarchical decomposition

$$p(\mathbf{z}_1, \dots, \mathbf{z}_L) = p(\mathbf{z}_L)p(\mathbf{z}_{L-1}|\mathbf{z}_L) \dots p(\mathbf{z}_1, \mathbf{z}_2);$$
$$q(\mathbf{z}_1, \dots, \mathbf{z}_L|\mathbf{x}) = q(\mathbf{z}_1|\mathbf{x}) \dots q(\mathbf{z}_L|\mathbf{x}).$$

ELBO

$$\begin{aligned}
\mathcal{L}(q, \boldsymbol{\theta}) &= \mathbb{E}_{q(\mathbf{z}_1|\mathbf{x})} \log p(\mathbf{x}|\mathbf{z}_1, \boldsymbol{\theta}) - KL(q(\mathbf{z}_1, \dots, \mathbf{z}_L|\mathbf{x})||p(\mathbf{z}_1, \dots, \mathbf{z}_L)) \\
&= \mathbb{E}_{q(\mathbf{z}_1|\mathbf{x})} \log p(\mathbf{x}|\mathbf{z}_1, \boldsymbol{\theta}) - \int \prod_{j=1}^{L} q(\mathbf{z}_j|\mathbf{x}) \sum_{i=1}^{L} \log \frac{q(\mathbf{z}_i|\mathbf{x})}{p(\mathbf{z}_i|\mathbf{z}_{i+1})} d\mathbf{z}_1 \dots d\mathbf{z}_L \\
&= \mathbb{E}_{q(\mathbf{z}_1|\mathbf{x})} \log p(\mathbf{x}|\mathbf{z}_1, \boldsymbol{\theta}) - \sum_{i=1}^{L} \int \prod_{j=1}^{L} q(\mathbf{z}_j|\mathbf{x}) \log \frac{q(\mathbf{z}_i|\mathbf{x})}{p(\mathbf{z}_i|\mathbf{z}_{i+1})} d\mathbf{z}_1 \dots d\mathbf{z}_L \\
&= \mathbb{E}_{q(\mathbf{z}_1|\mathbf{x})} \log p(\mathbf{x}|\mathbf{z}_1, \boldsymbol{\theta}) - \sum_{i=1}^{L} \int q(\mathbf{z}_{i+1}|\mathbf{x}) q(\mathbf{z}_i|\mathbf{x}) \log \frac{q(\mathbf{z}_i|\mathbf{x})}{p(\mathbf{z}_i|\mathbf{z}_{i+1})} d\mathbf{z}_i d\mathbf{z}_{i+1} \\
&= \mathbb{E}_{q(\mathbf{z}_1|\mathbf{x})} \log p(\mathbf{x}|\mathbf{z}_1, \boldsymbol{\theta}) - \sum_{i=1}^{L} \mathbb{E}_{q(\mathbf{z}_{i+1}|\mathbf{x})} \left[KL(q(\mathbf{z}_i|\mathbf{x})||p(\mathbf{z}_i|\mathbf{z}_{i+1})) \right]
\end{aligned}$$

---

https://arxiv.org/pdf/1611.05013.pdf

PixelVAE, 2016

$$p(\mathbf{x}|\boldsymbol{\theta}) = \int p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) d\mathbf{z} = \int p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) p(\mathbf{z}) d\mathbf{z}$$

How to make the generative model $p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta})$ more powerful?

Autoregressive decoder

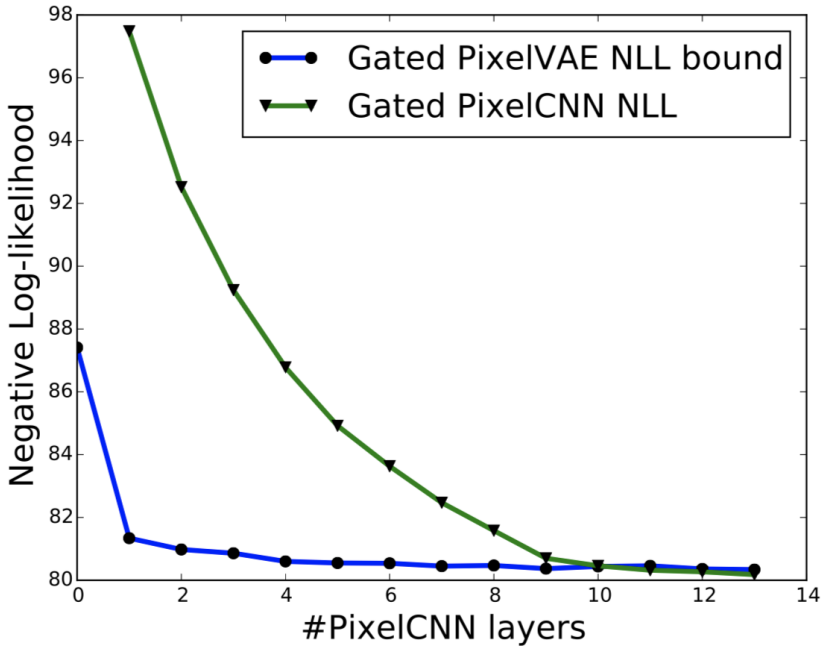
$$p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) = \prod_{i=1}^{n} p(x_i|\mathbf{x}_{1:i-1}, \mathbf{z}, \boldsymbol{\theta})$$

Good idea for density estimation, but could be harmful for learning good representations!

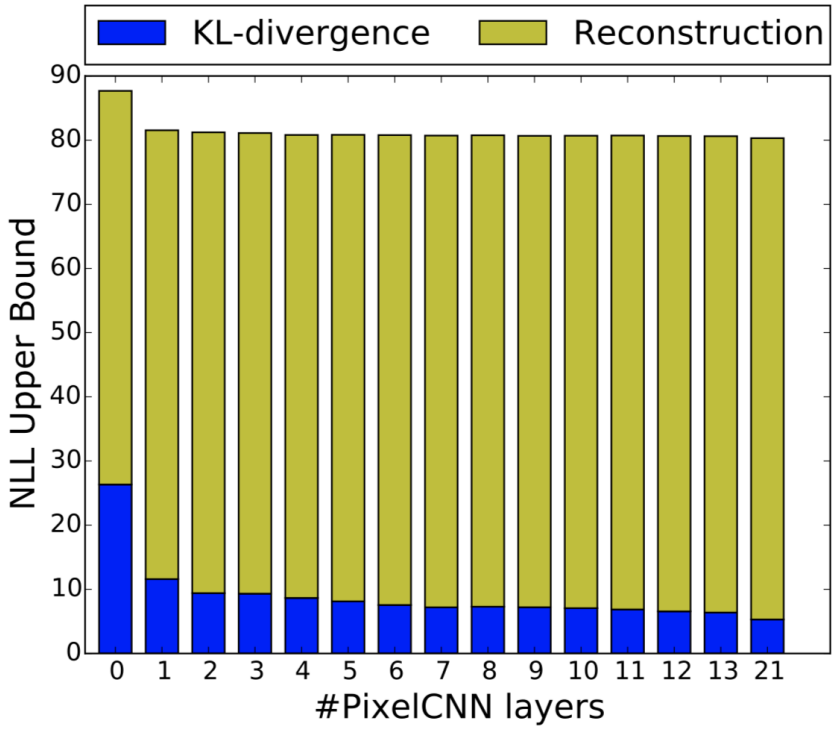
---

https://arxiv.org/pdf/1611.05013.pdf

PixelVAE, 2016



(a)



(b)

---

https://arxiv.org/pdf/1611.05013.pdf

PixelVAE, 2016

https://arxiv.org/pdf/1611.05013.pdf

PixelVAE, 2016

Model	NLL Validation (Train)
Convolutional DRAW (Gregor et al., 2016)	$\leq$ 4.10 (4.04)
Real NVP (Dinh et al., 2016)	$=$ 4.01 (3.93)
PixelRNN (van den Oord et al., 2016a)	$=$ 3.63 (3.57)
Gated PixelCNN (van den Oord et al., 2016b)	$=$ **3.57** (3.48)
Hierarchical PixelVAE	$\leq$ 3.66 (3.59)

https://arxiv.org/pdf/1611.05013.pdf

Variational Lossy Autoencoder, 2016

Representation learning

"Identifies and disentangles the underlying causal factors of the data, so that it becomes easier to understand the data, to classify it, or to perform other tasks".

$$p(\mathbf{x}|\boldsymbol{\theta}) = \int p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta}) d\mathbf{z} = \int p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) p(\mathbf{z}) d\mathbf{z}$$

Posterior collapse

Let consider extreme case, where $p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta})$ could be any distribution.
Then optimum will be at $p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) = \pi(\mathbf{x})$.

---

https://arxiv.org/pdf/1611.02731.pdf

Decoder weakening

How to force the model encode information about $\mathbf{x}$ into $\mathbf{z}$?

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

What we get if $\beta = 1$ ($\beta = 0$)?

KL annealing

- Start training with $\beta = 0$.
- Increase it until $\beta = 1$ during training process.

---

https://arxiv.org/abs/1511.06349

Decoder weakening

Free bits

- Divide the latent dimensions into the K subsets.
- Ensure that using less than λ nats of information per subset j.

$$\hat{\mathcal{L}}(q, \theta) = \mathbb{E}_{q(\mathbf{Z}|\mathbf{X})} \log p(\mathbf{X}|\mathbf{Z}, \theta) - \sum_{j=1}^{K} \max(\lambda, KL(q(\mathbf{Z}_j|\mathbf{X})||p(\mathbf{Z}_j))).$$

Increasing the latent information is advantageous for the reconstruction term.

This results in $KL(q(\mathbf{Z}_j|\mathbf{x})||p(\mathbf{Z}_j)) \geq \lambda$ for all j, in practice.

---

https://arxiv.org/abs/1606.04934

Variational Lossy AutoEncoder, 2016

Lossy code via explicit information placement

$$p(\mathbf{x}|\mathbf{z},\theta) = \prod_{i=1}^{m} p\big(x_i|\mathbf{z}, \mathbf{x}_{\text{WindowAround}(i)}\big).$$

- WindowAround(i) restricts the receptive field (it forbids to represent arbitrarily complex distribution over $\mathbf{x}$ without dependence on $\mathbf{z}$).
- Local statistics of 2D images (texture) will be modeled by a small local window.
- Global structural information (shapes) is long-range dependency that can only be communicated through latent code $\mathbf{z}$.

---

https://arxiv.org/abs/1606.04934

Variational Lossy AutoEncoder, 2016

Theorem

AF prior is equivalent to IAF posterior.

Proof

$$\mathcal{L}(q, \theta) = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \left[\log p(\mathbf{x}|\mathbf{z}, \theta) + \log p(\mathbf{z}) - q(\mathbf{z}|\mathbf{x})\right]$$

$$= \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}|\mathbf{x}), \mathbf{z}_0 = f(\mathbf{z})} \Big[\log p(\mathbf{x}|g(\mathbf{z}_0), \theta) + \underbrace{\left(\log p(\mathbf{z}_0) + \log \left|\det \frac{\partial \mathbf{z}_0}{\partial \mathbf{z}}\right|\right)}_{\text{AF prior}} - q(\mathbf{z}|\mathbf{x})\Big]$$

$$= \mathbb{E}_{\mathbf{z}\sim q(\mathbf{z}|\mathbf{x}), \mathbf{z}_0 = f(\mathbf{z})} \Big[\log p(\mathbf{x}|g(\mathbf{z}_0), \theta) + \log p(\mathbf{z}_0) - \underbrace{\left(q(\mathbf{z}|\mathbf{x}) - \log \left|\det \frac{\partial \mathbf{z}_0}{\partial \mathbf{z}}\right|\right)}_{\text{IAF posterior}}\Big]$$

- AF prior is the same as IAF posterior along the encoder path, $f(q(\mathbf{z}|\mathbf{x}))$,
- IAF posterior has a shorter decoder path $p(\mathbf{x}|\mathbf{z})$, AF prior has a deeper decoder path $p(\mathbf{x}|g(\mathbf{z}_0))$.

AF prior and IAF posterior have the same computation cost, so using AF prior makes the model more expressive at no training time cost.

---

https://arxiv.org/abs/1606.04934

Variational Lossy AutoEncoder, 2016

- Can VLAE learn lossy codes that encode global statistics?
- Does using AF priors improves upon using IAF posteriors as predicted by theory?
- Does using autoregressive decoding distributions improve density estimation performance?

Model	**NLL Test**
Normalizing flows (Rezende & Mohamed, 2015)	85.10
DRAW (Gregor et al., 2015)	< 80.97
Discrete VAE (Rolfe, 2016)	81.01
PixelRNN (van den Oord et al., 2016a)	79.20
IAF VAE (Kingma et al., 2016)	79.88
AF VAE	79.30
VLAE	**79.03**

Method	bits/dim $\leq$
Results with tractable likelihood models:	
Uniform distribution [1]	8.00
Multivariate Gaussian [1]	4.70
NICE [2]	4.48
Deep GMMs [3]	4.00
Real NVP [4]	3.49
PixelCNN [1]	3.14
Gated PixelCNN [5]	3.03
PixelRNN [1]	3.00
PixelCNN++ [6]	**2.92**
Results with variationally trained latent-variable models:	
Deep Diffusion [7]	5.40
Convolutional DRAW [8]	3.58
ResNet VAE with IAF [9]	3.11
ResNet VLAE	3.04
DenseNet VLAE	**2.95**

https://arxiv.org/abs/1606.04934

Disentangled representations

Goal

Learning an interpretable factorised representation of the independent data gen- erative 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

- $p(\mathbf{x}|\mathbf{v},\mathbf{w}) = \text{Sim}(\mathbf{v},\mathbf{w})$ – 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

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.

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

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

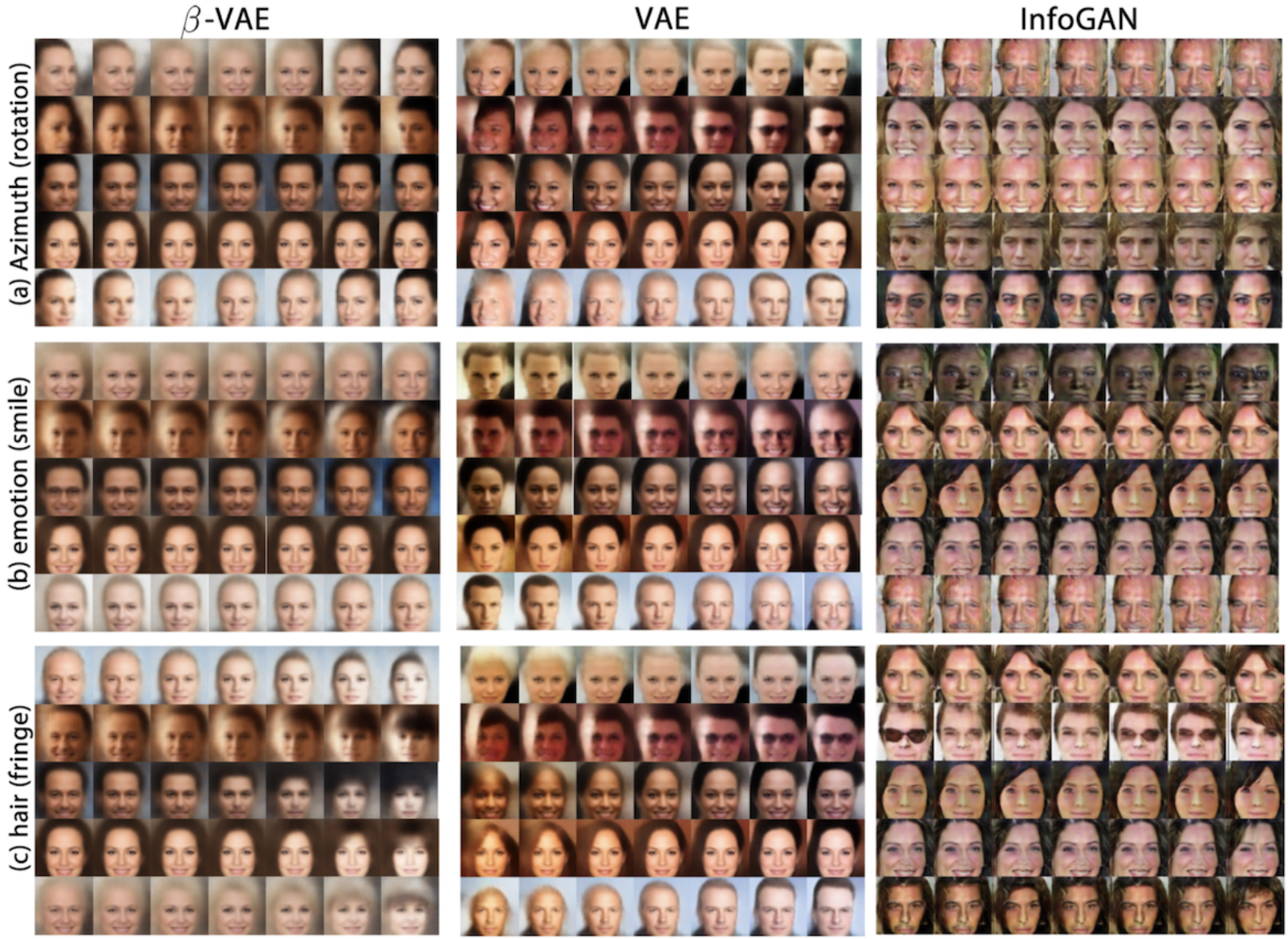
$$q(\mathbf{z}|\mathbf{x}) = \mathcal{N}\left(\mu(\mathbf{x})|\sigma^2(\mathbf{x})\right); \quad \mathbf{z}_{li} = \mu(\mathbf{x}_{li}); \quad \mathbf{z}_{lj} = \mu(\mathbf{x}_{lj}).$$

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

---

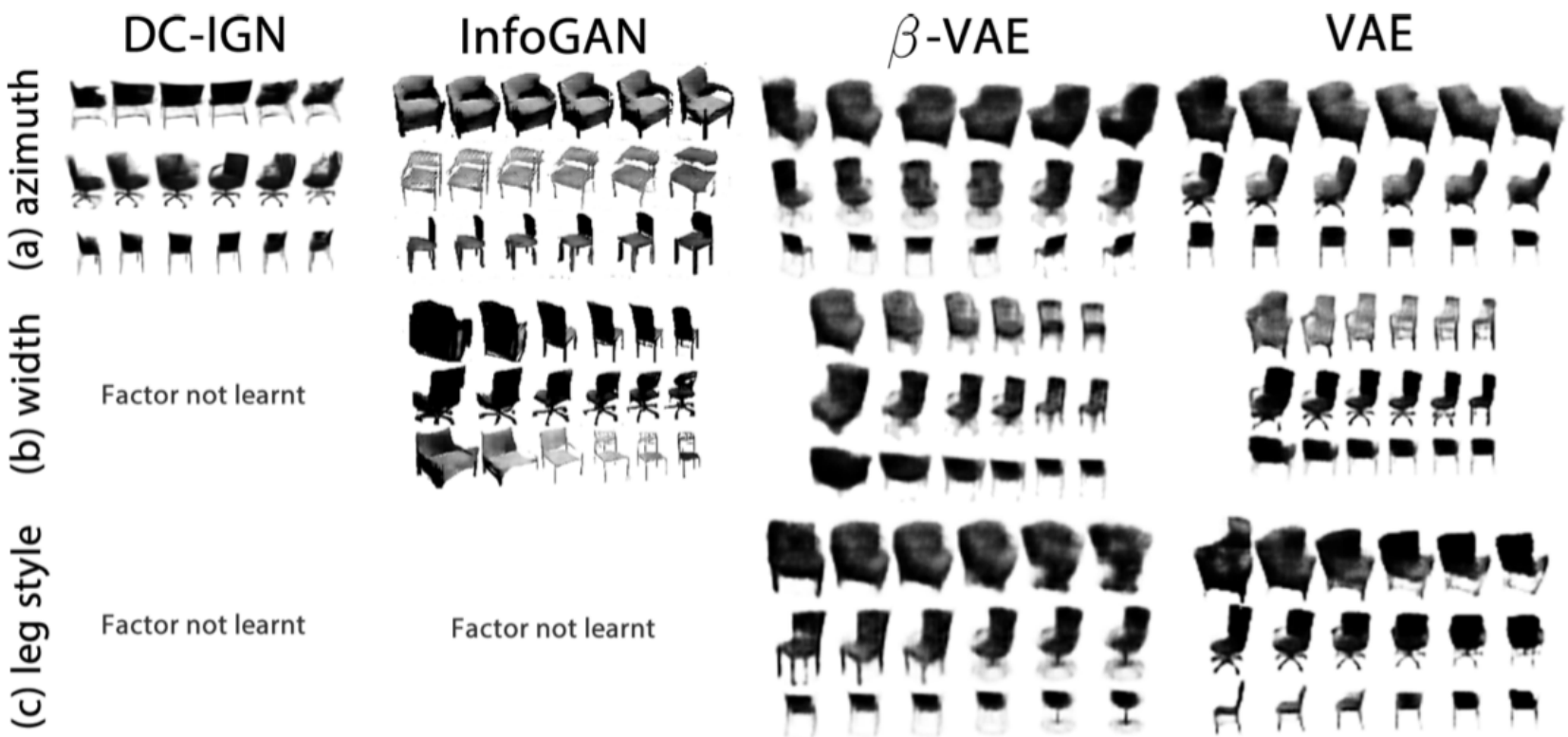
https://openreview.net/references/pdf?id=Sy2fzU9gl

β-VAE, 2017



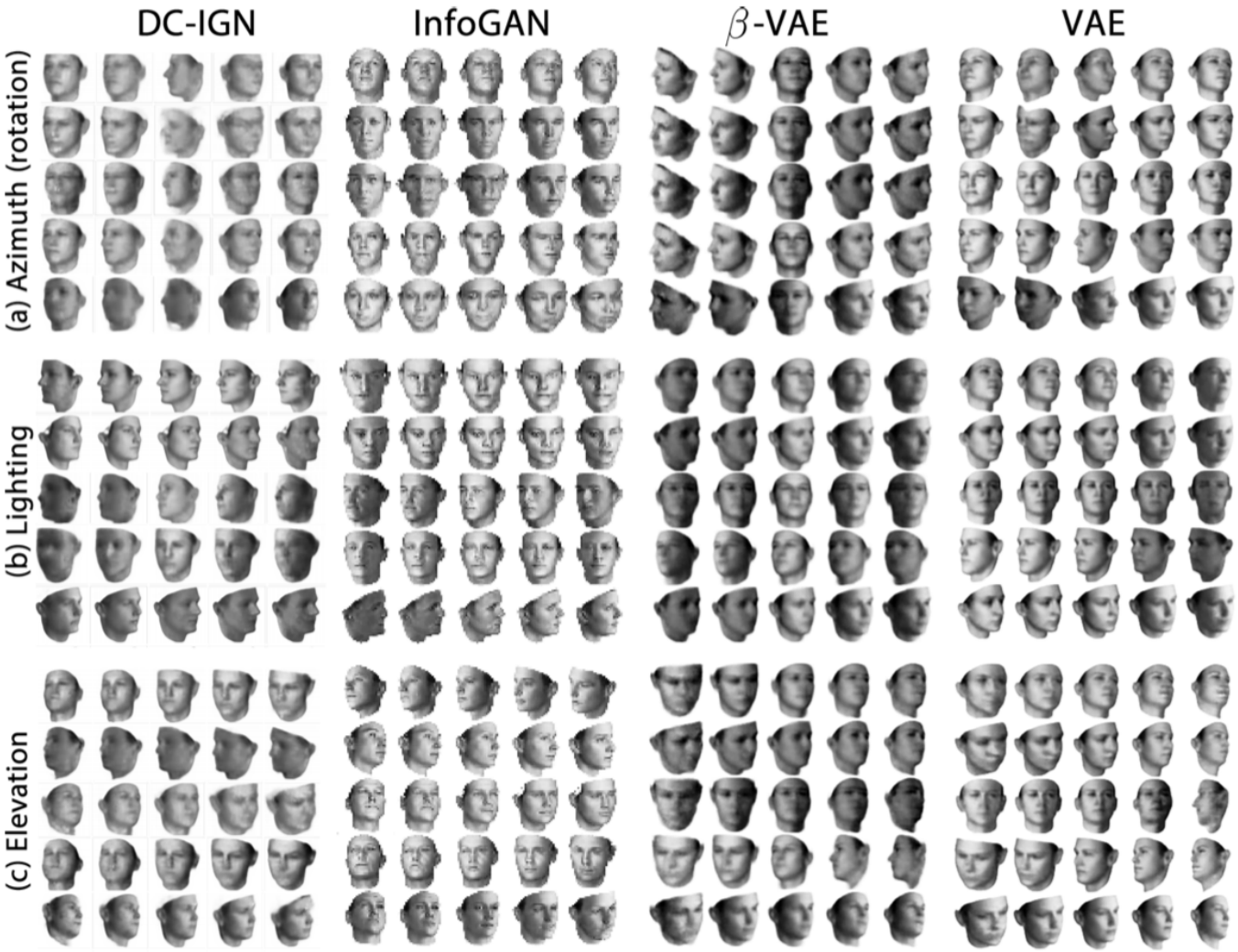
https://openreview.net/references/pdf?id=Sy2fzU9gl

β-VAE, 2017



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

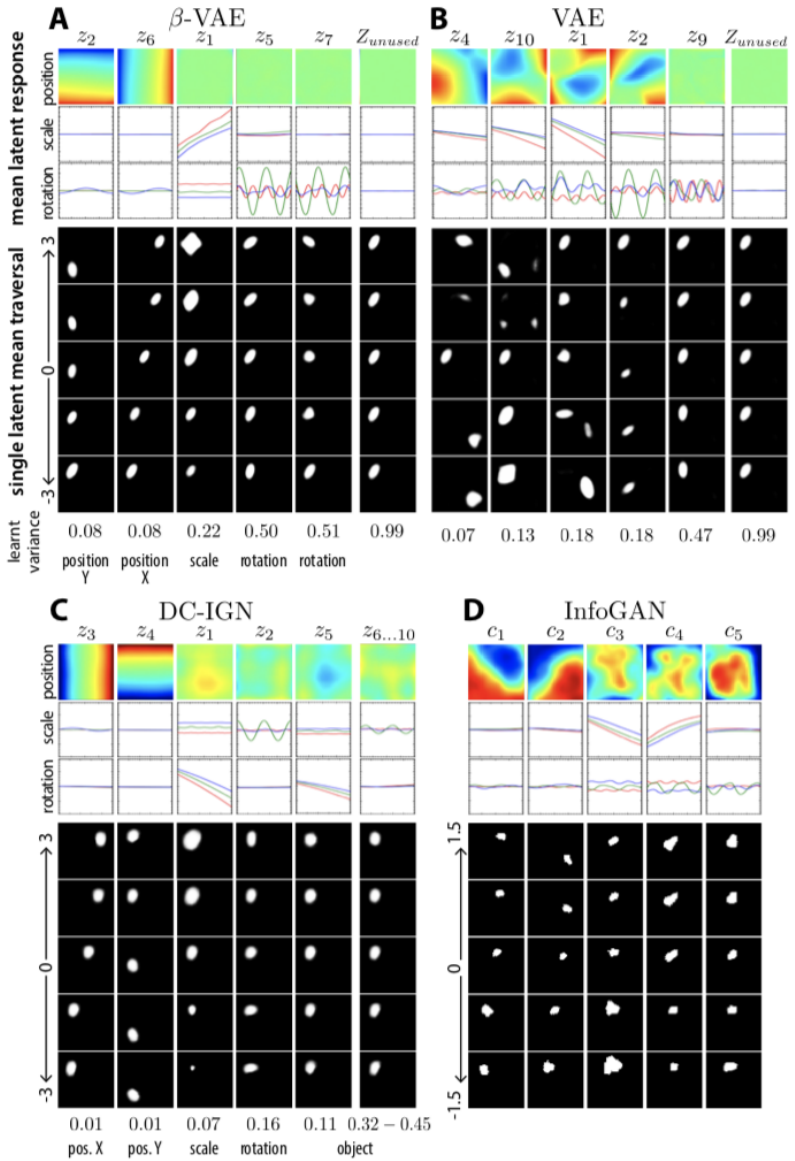
β-VAE, 2017



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

β-VAE, 2017

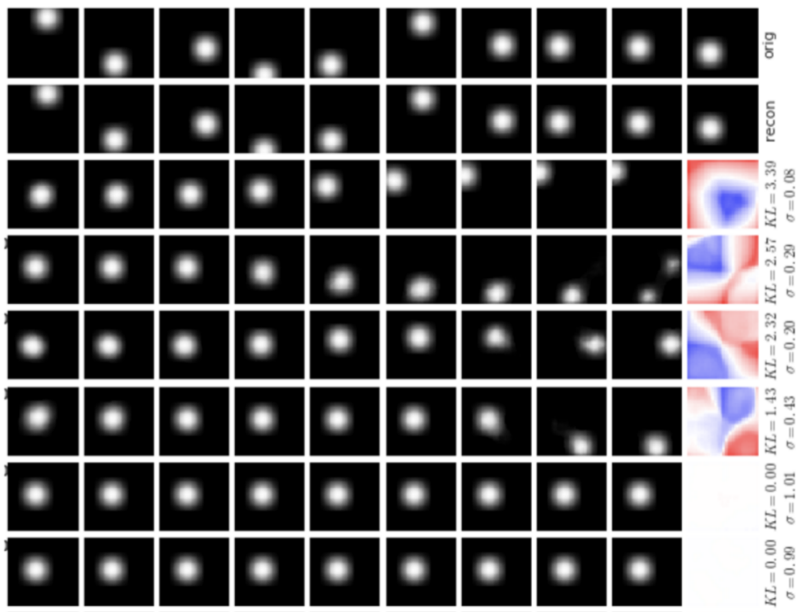
Model	Disentanglement metric score
Ground truth	*100%*
Raw pixels	$45.75 \pm 0.8\%$
PCA	$84.9 \pm 0.4\%$
ICA	$42.03 \pm 10.6\%$
DC-IGN	$\mathbf{99.3 \pm 0.1\%}$
InfoGAN	$73.5 \pm 0.9\%$
VAE untrained	$44.14 \pm 2.5\%$
VAE	$61.58 \pm 0.5\%$
β-VAE	$\mathbf{99.23 \pm 0.1\%}$



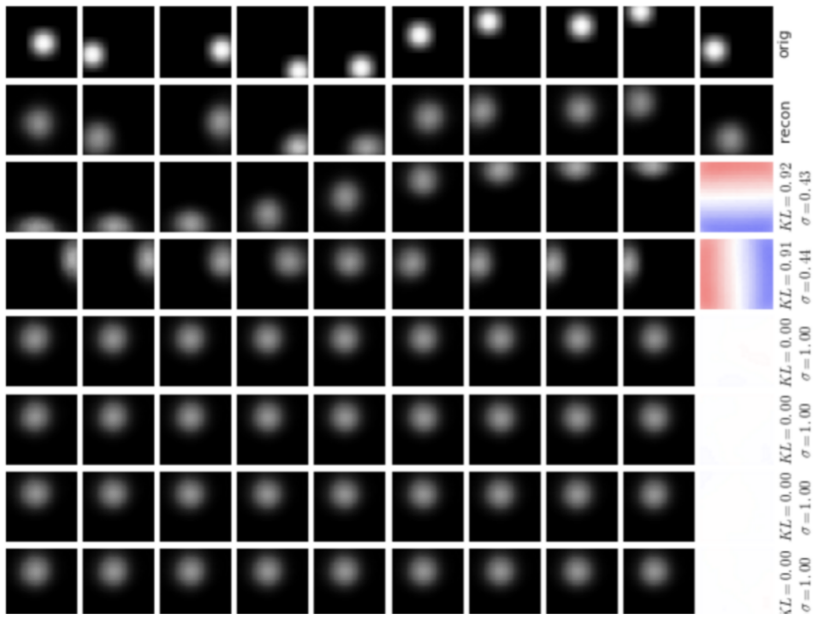
https://openreview.net/references/pdf?id=Sy2fzU9gl

β-VAE, 2018

β = 1



β = 150



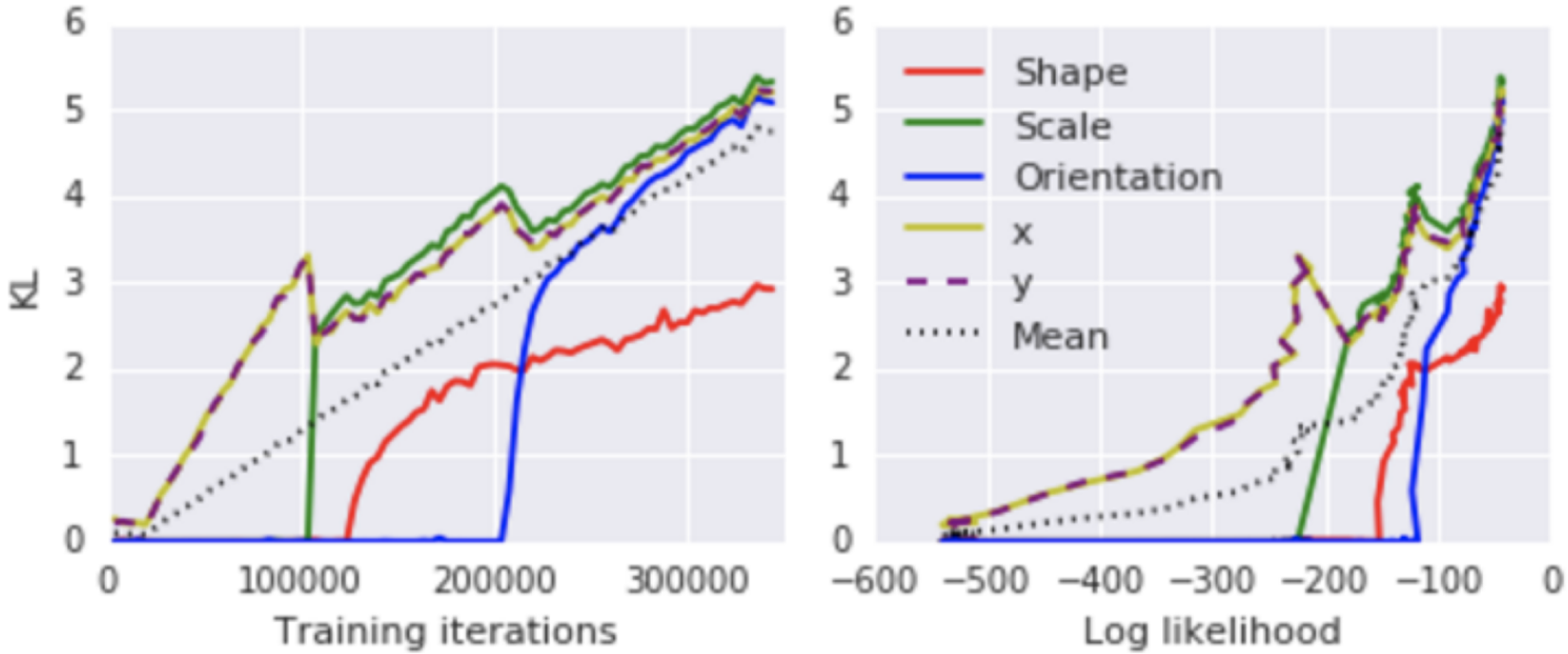
---

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

β-VAE, 2018

Controlled encoding capacity

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

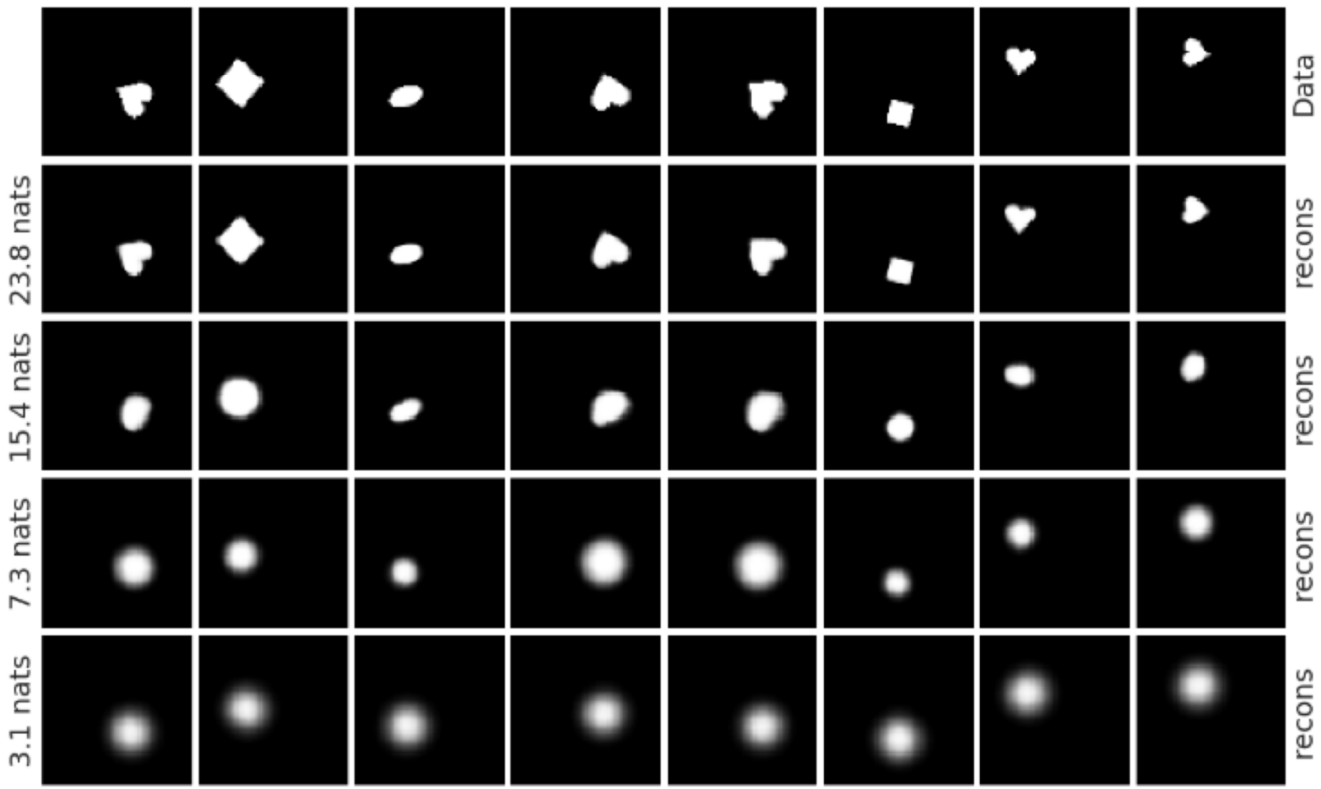


https://arxiv.org/pdf/1804.03599.pdf

β-VAE, 2018

Controlled encoding capacity

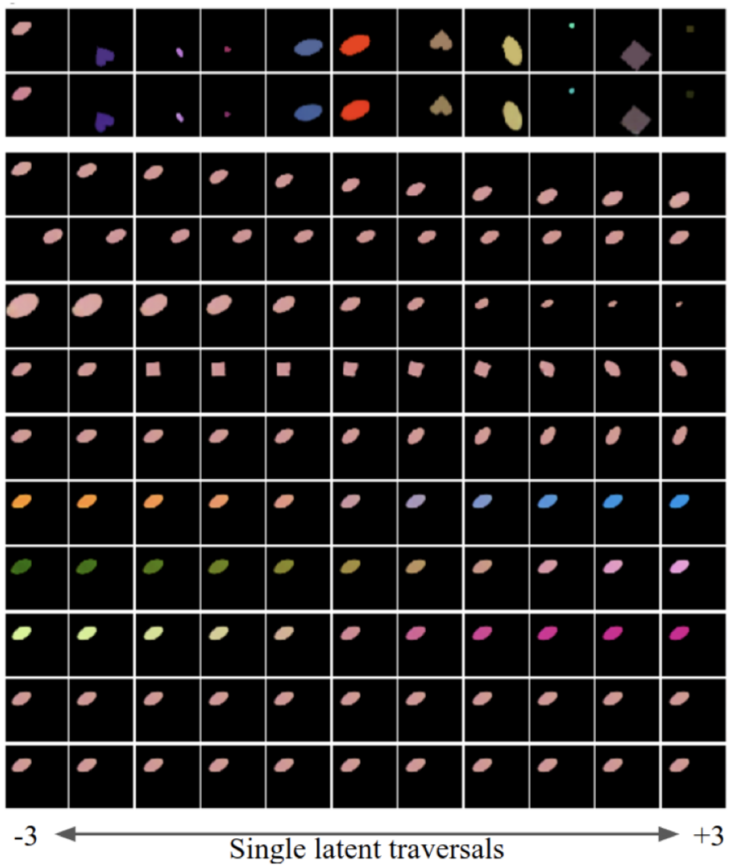
$$\mathcal{L}(q, \boldsymbol{\theta}, \beta) = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \log p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) - |KL(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) - C|.$$



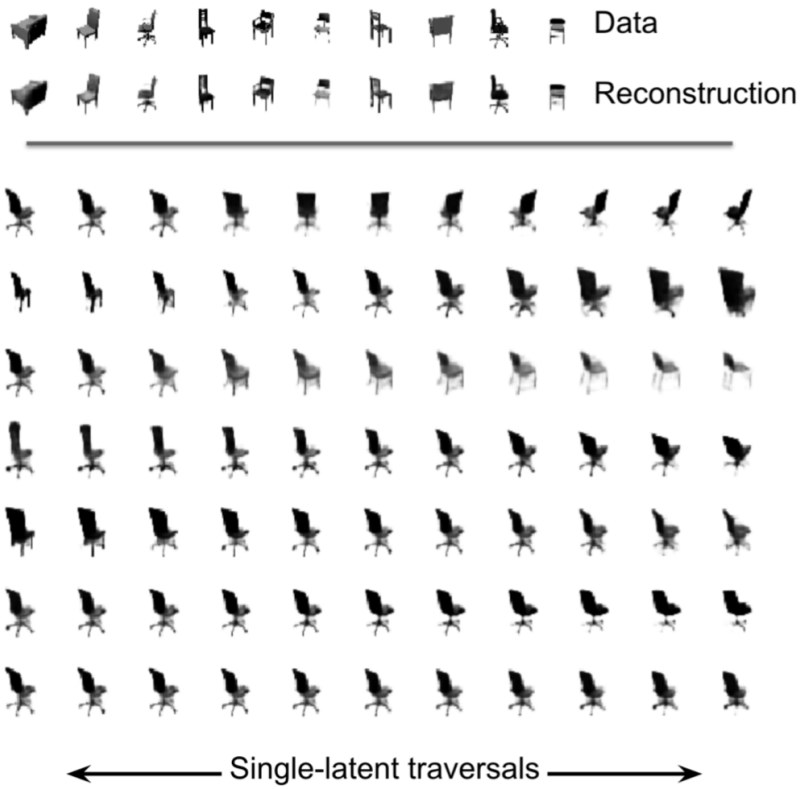
https://arxiv.org/pdf/1804.03599.pdf

β-VAE, 2018

(a) Coloured dSprites



(b) 3D Chairs



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

References

- **PixelVAE:** A Latent Variable Model for Natural Images
https://arxiv.org/pdf/1611.05013.pdf
Summary: Use autoregressive model (PixelCNN) in VAE decoder. Use hierarchical structure of latent variables. Restrict the number of conv layers in PixelCNN (receptive field) to encode global structure in latent space. The performance is slightly worse than GatedPixelCNN.
- Generating Sentences from a Continuous Space
https://arxiv.org/abs/1511.06349
Summary: KL annealing proposed to weaken autoregressive decoder.
- Improving Variational Inference with Inverse Autoregressive Flow
https://arxiv.org/abs/1606.04934
Summary: Free bits proposed to weaken autoregressive decoder.
- **VLAE:** Variational Lossy Autoencoder
https://arxiv.org/abs/1611.02731
Summary: Bits-back coding interpretation for posterior collapse. Solving the problem of ignoring latent codes. Reduce receptive field in PixelCNN decoder to encode global information, use learnable AF prior.
- **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.