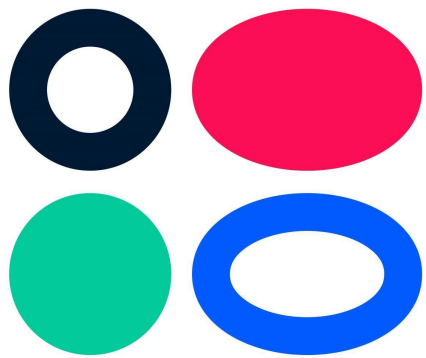
Deep Generative Models
Lecture 8

Roman Isachenko



Ozon Masters

Spring, 2021

ELBO surgery

Theorem

$$\frac{1}{n}\sum_{i=1}^{n} KL(q(\mathbf{z}|\mathbf{x}_i)||p(\mathbf{z})) = KL(q(\mathbf{z})||p(\mathbf{z})) + \mathbb{I}_q[\mathbf{x},\mathbf{z}].$$

ELBO revisiting

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

$$= \underbrace{\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{q(\mathbf{z}|\mathbf{x}_i)}\log p(\mathbf{x}_i|\mathbf{z},\boldsymbol{\theta})}_{\text{Reconstruction loss}} - \underbrace{\mathbb{I}_q[\mathbf{x},\mathbf{z}]}_{0\leq \text{MI} \leq \log n} \underbrace{- KL(q(\mathbf{z})||p(\mathbf{z}))}_{\text{Marginal KL}}$$

Prior distribution $p(\mathbf{z})$ is only in the last term.

Hoffman M. D., Johnson M. J. ELBO surgery: yet another way to carve up the variational evidence lower bound, 2016

ELBO surgery

ELBO revisiting

$$\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_i(q,\theta) = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{q(\mathbf{z}|\mathbf{x}_i)}\log p(\mathbf{x}_i|\mathbf{z},\theta)}_{\text{Reconstruction loss}} - \underbrace{\mathbb{I}_q[\mathbf{x},\mathbf{z}]}_{0\leq \text{MI}\leq \log n} - \underbrace{KL(q(\mathbf{z})||p(\mathbf{z}))}_{\text{Marginal KL}}$$

$$KL(q(\mathbf{z})||p(\mathbf{z})) = 0 \quad \Leftrightarrow \quad p(\mathbf{z}) = q(\mathbf{z}) = \frac{1}{n}\sum_{i=1}^{n} q(\mathbf{z}|\mathbf{x}_i).$$

The optimal prior distribution $p(\mathbf{z})$ is aggregated posterior $q(\mathbf{z})$.

	ELBO	Avg. KL	Mutual info. ②	Marg. KL ③
2D latents	-129.63	7.41	7.20	0.21
10D latents	-88.95	19.17	10.82	8.35
20D latents	-87.45	20.2	10.67	9.53

$$\log n \approx 11.0021$$

Hoffman M. D., Johnson M. J. ELBO surgery: yet another way to carve up the variational evidence lower bound, 2016

VAE prior

ELBO revisiting

$$\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_i(q,\boldsymbol{\theta}) = \underbrace{\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{q(\mathbf{z}|\mathbf{x}_i)}\log p(\mathbf{x}_i|\mathbf{z},\boldsymbol{\theta})}_{\text{Reconstruction loss}} - \underbrace{\mathbb{I}_q[\mathbf{x},\mathbf{z}]}_{0\leq \text{MI} \leq \log n} \underbrace{- KL(q(\mathbf{z})||p(\mathbf{z}))}_{\text{Marginal KL}}$$

How to choose the optimal $p(\mathbf{z})$?

- Standard Gaussian: $p(\mathbf{z}) = \mathcal{N}(0, I) \Rightarrow$ over-regularization;
- Mixture of Gaussians: $p(\mathbf{z}|\boldsymbol{\lambda}) = \frac{1}{K}\sum_{k=1}^{K}\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\sigma}_k^2)$ ([1], [2]);
- $p(\mathbf{z}) = q(\mathbf{z}) = \frac{1}{n}\sum_{i=1}^{n} q(\mathbf{z}|\mathbf{x}_i) \Rightarrow$ overfitting and highly expensive.

Note that now we consider cases of parametrized priors.

---

[1] Dilokthanakul N. et al. Deep Unsupervised Clustering with Gaussian Mixture Variational Autoencoders, 2016

[2] Nalisnick E., Hertel L., Smyth P. Approximate Inference for Deep Latent Gaussian Mixtures, 2016

VampPrior

Variational Mixture of posteriors

$$p(\mathbf{z}|\boldsymbol{\lambda}) = \frac{1}{K}\sum_{k=1}^{K} q(\mathbf{z}|\mathbf{u}_k),$$

where $\boldsymbol{\lambda} = \{\mathbf{u}_1, \ldots, \mathbf{u}_K\}$ are trainable preudo-inputs.

- Multimodal $\Rightarrow$ prevents over-regularization;.
- $K \ll n \Rightarrow$ prevents from potential overfitting + less expensive to train.
- Pseudo-inputs are prior hyperparameters $\Rightarrow$ connection to the Empirical Bayes.

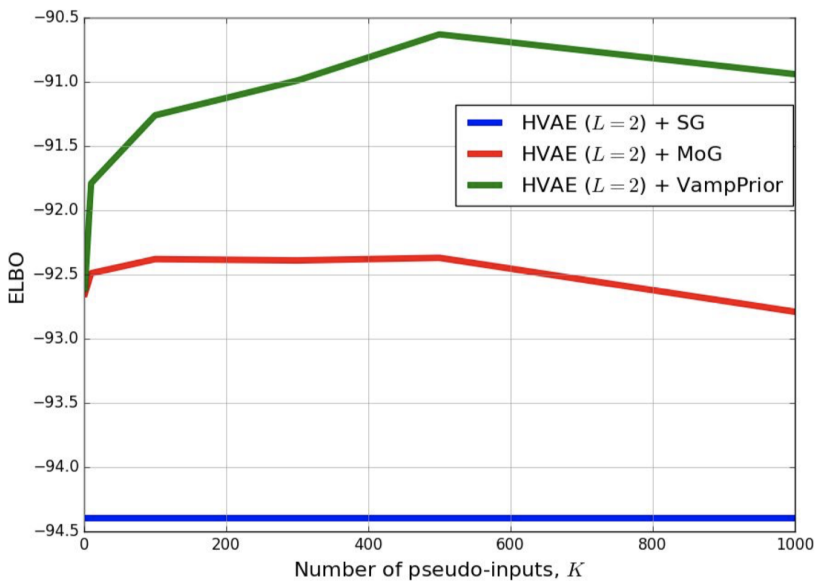
Tomczak J. M., Welling M. VAE with a VampPrior, 2017

VampPrior

- Do we equally need the multimodal prior?
- Is it beneficial to couple the prior with the variational posterior or MoG is enough?

MNIST results

Model	LL
VAE ($L=1$) + NF [32]	-85.10
VAE ($L=2$) [6]	-87.86
IWAE ($L=2$) [6]	-85.32
HVAE ($L=2$) + SG	-85.89
HVAE ($L=2$) + MoG	-85.07
HVAE ($L=2$) + VampPrior *data*	-85.71
HVAE ($L=2$) + VampPrior	$\mathbf{-83.19}$
AVB + AC ($L=1$) [28]	-80.20
VLAE [7]	$\mathbf{-79.03}$
VAE + IAF [18]	-79.88
convHVAE ($L=2$) + VampPrior	-81.09
PixelHVAE ($L=2$) + VampPrior	-79.78

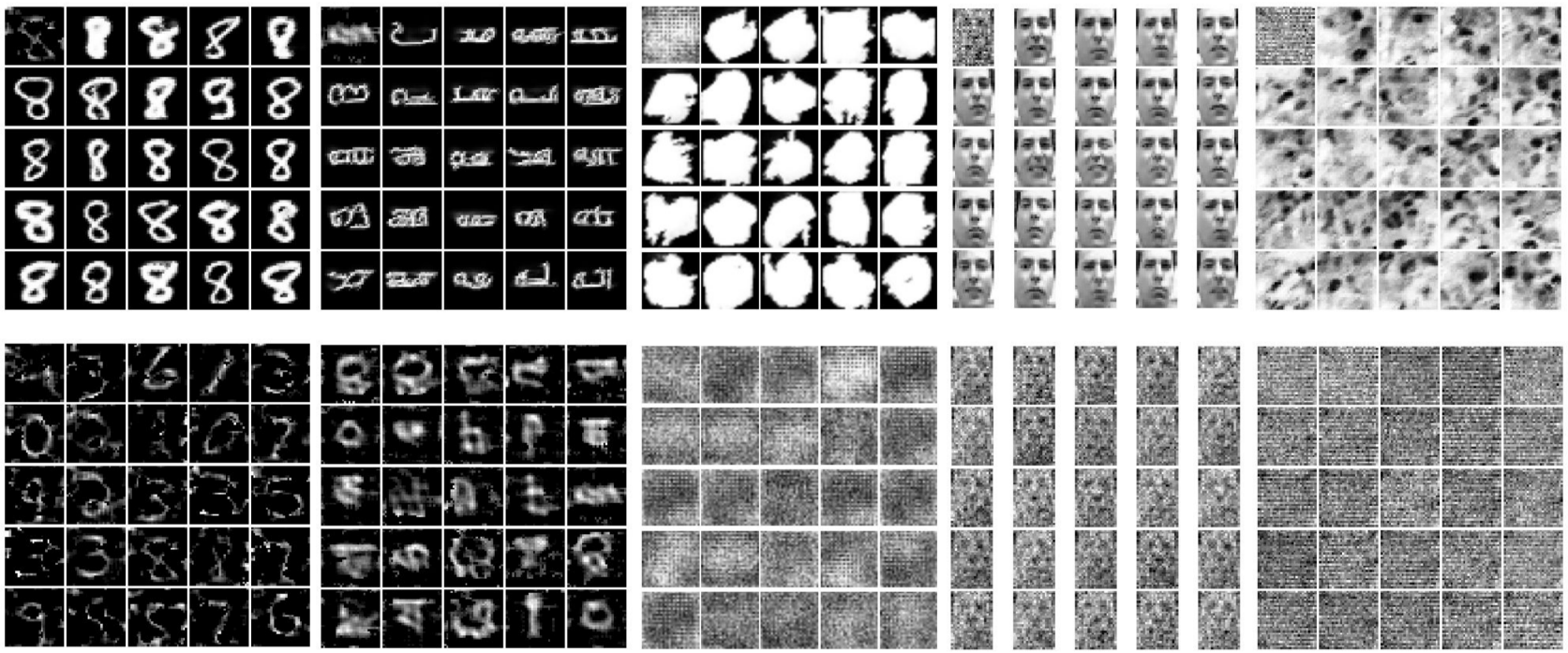


Tomczak J. M., Welling M. VAE with a VampPrior, 2017

VampPrior

Top row: generated images by PixelHVAE + VampPrior for chosen pseudo-input in the left top corner.

Bottom row: pseudo-inputs for different datasets.



Tomczak J. M., Welling M. VAE with a VampPrior, 2017

Flow prior in VAE

ELBO revisiting

$$\mathcal{L}(q, \theta) = \underbrace{\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{q(\mathbf{z}_i|\mathbf{x}_i)} \log p(\mathbf{x}_i|\mathbf{z}_i, \theta)}_{\text{Reconstruction loss}} - \underbrace{\left(\log n - \mathbb{E}_{q(\mathbf{z})}\mathbb{H}\left[q(i|\mathbf{z})\right]\right)}_{0 \leq \text{Mutual info} \leq \log N} - \underbrace{KL(q(\mathbf{z})||p(\mathbf{z}))}_{\text{Marginal KL}}$$

VampPrior

$$p(\mathbf{z}|\boldsymbol{\lambda}) = \frac{1}{K}\sum_{k=1}^{K} q(\mathbf{z}|\mathbf{u}_k),$$

where $\mathbf{u}_1, \dots \mathbf{u}_K$ is trainable preudo-inputs.

Autoregressive flow prior

$$\log p(\mathbf{z}|\boldsymbol{\lambda}) = \log p(\boldsymbol{\epsilon}) + \log \det \left|\frac{d\boldsymbol{\epsilon}}{d\mathbf{z}}\right|$$

$$\mathbf{z} = g(\boldsymbol{\epsilon}, \boldsymbol{\lambda}) = f^{-1}(\boldsymbol{\epsilon}, \boldsymbol{\lambda})$$

Chen X. et al. Variational Lossy Autoencoder, 2016

Flow prior in VAE

Theorem

VAE with the AF prior for latent code $\mathbf{z}$ is equivalent to using the IAF posterior for latent code $\boldsymbol{\epsilon}$.

Proof

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

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

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

Flows in VAE

$$\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta}) = \mathbb{E}_{q(\mathbf{z}_0|\mathbf{x}, \boldsymbol{\phi})} \left[\log p(\mathbf{x}, \mathbf{z}_K|\boldsymbol{\theta}) - \log q(\mathbf{z}_0|\mathbf{x}, \boldsymbol{\phi}) + \log \left|\det \left(\frac{\partial g(\mathbf{z}_0, \boldsymbol{\phi}_*)}{\partial \mathbf{z}_0}\right)\right|\right].$$

Chen X. et al. Variational Lossy Autoencoder, 2016

Flow prior in VAE

Autoregressive flow prior

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

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

- IAF posterior decoder path: $p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta})$, $\mathbf{z} \sim p(\mathbf{z})$.
- AF prior decoder path: $p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta})$, $\mathbf{z} = g(\boldsymbol{\epsilon}, \boldsymbol{\lambda})$, $\boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon})$.

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

Chen X. et al. Variational Lossy Autoencoder, 2016

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)

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

- Loose lower bound

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

Posterior collapse: toy example

Let define latent variable model in the following way:

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

- prior distribution $p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|0, \mathbf{I})$;
- probabilistic model $p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathbf{z}), \boldsymbol{\sigma}_{\boldsymbol{\theta}}(\mathbf{z}))$ (diagonal covariance);
- variational posterior $q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{x}), \boldsymbol{\sigma}_{\phi}(\mathbf{x}))$ (diagonal covariance).

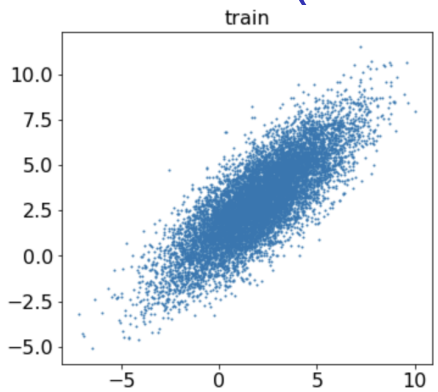
Let data distribution is $\pi(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Possible cases:

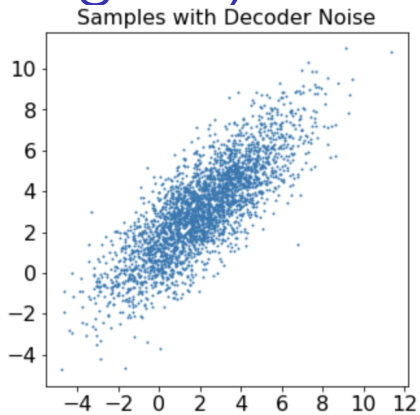
- covariance matrix $\boldsymbol{\Sigma}$ is diagonal (univariate case);
- covariance matrix $\boldsymbol{\Sigma}$ is **not** diagonal (multivariate case).

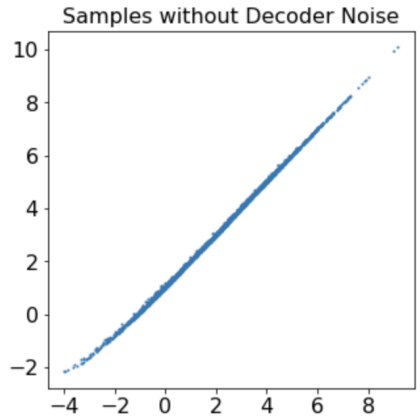
What is the difference?

Posterior collapse: toy example

Multivariate ($\mathbf{\Sigma}$ is non-diagonal)

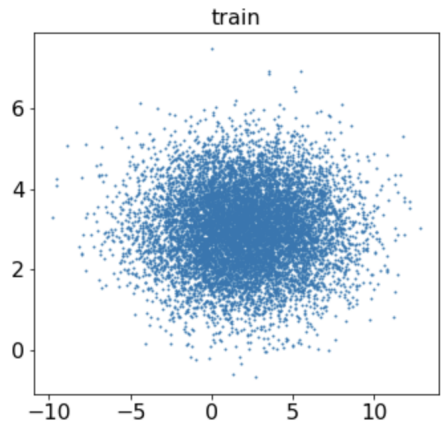


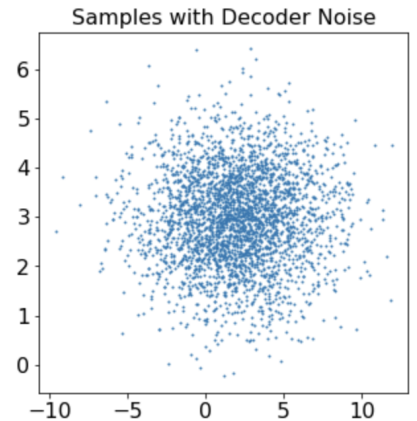


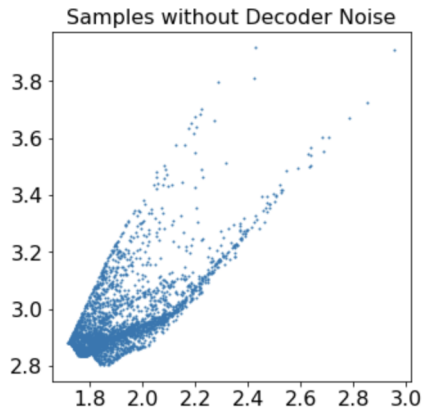


The encoder uses latent variables to model data.

Univariate ($\mathbf{\Sigma}$ is diagonal)







Latent variables are not used, since the decoder could model the data without the encoder.

Posterior collapse

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".

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

If the decoder model $p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta})$ is powerful enough to model $p(\mathbf{x}|\boldsymbol{\theta})$ the latent variables $\mathbf{z}$ becames irrelevant.

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

Early in the training the approximate posterior $q(\mathbf{z}|\mathbf{x})$ carries little information about $\mathbf{x}$ and the model sets the posterior to the prior to avoid paying any cost $KL(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))$.

PixelVAE

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

- More powerful $p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta})$ leads to more powerful generative model $p(\mathbf{x}|\boldsymbol{\theta})$.
- Too powerful $p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta})$ could lead to posterior collapse, where variational posterior $q(\mathbf{z}|\mathbf{x})$ will not carry any information about data and close to prior $p(\mathbf{z})$.

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

Autoregressive decoder

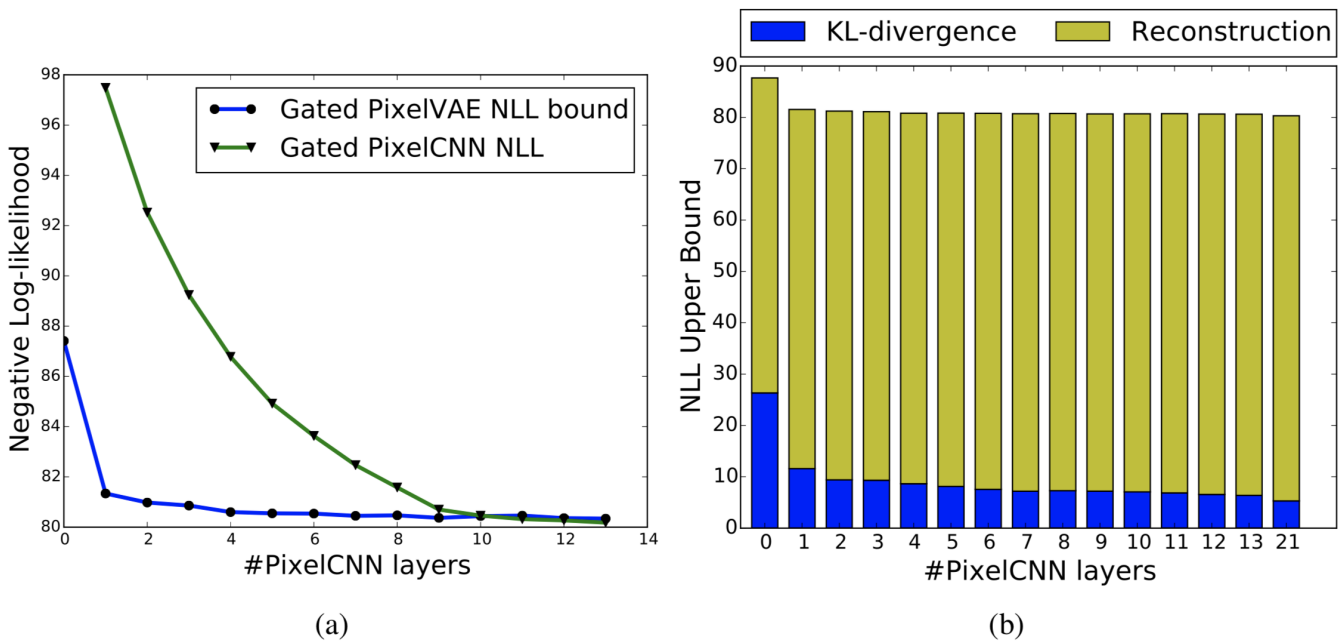
$$p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) = \prod_{i=1}^{n} p(x_i|\mathbf{x}_{1:i-1}, \mathbf{z}, \boldsymbol{\theta})$$

Gulrajani I. et al. PixelVAE: A Latent Variable Model for Natural Images, 2016

PixelVAE

VAE model with autoregressive PixelCNN decoder with few autoregressive layers.

- Global structure is captured by latent variables.
- Local statistics are captured by limited receptive field autoregressive model.



(a) (b)

Gulrajani I. et al. PixelVAE: A Latent Variable Model for Natural Images, 2016

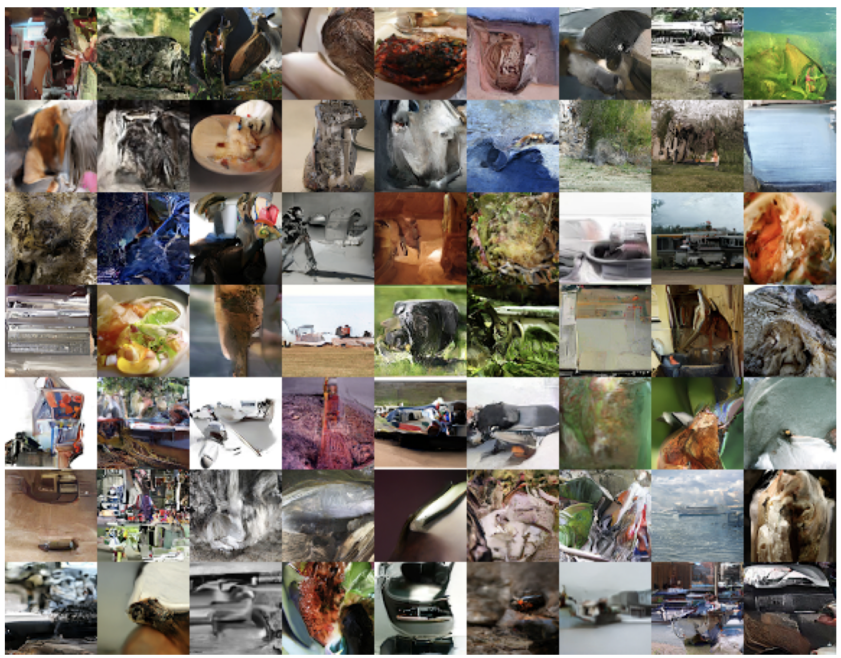
PixelVAE

MNIST

Model	NLL Test
DRAW (Gregor et al., 2016)	$\leq$ 80.97
Discrete VAE (Rolfe, 2016)	= 81.01
IAF VAE (Kingma et al., 2016)	$\approx$ 79.88
PixelCNN (van den Oord et al., 2016a)	= 81.30
PixelRNN (van den Oord et al., 2016a)	= 79.20
Convolutional VAE	$\leq$ 87.41
PixelVAE	$\leq$ 80.64
Gated PixelCNN (our implementation)	= 80.10
Gated PixelVAE	$\approx$ 79.48 ($\leq$ 80.02)
Gated PixelVAE without upsampling	$\approx$ **79.02** ($\leq$ 79.66)

ImageNet 64x64

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)



Gulrajani I. et al. PixelVAE: A Latent Variable Model for Natural Images, 2016

Decoder weakening

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

Powerful decoder $p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta})$ makes the model expressive, but posterior collapse is possible.
PixelVAE model uses the autoregressive PixelCNN model with small number of layers to limit receptive field.
How to force the model encode information about $\mathbf{x}$ into $\mathbf{z}$?

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

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

KL annealing

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

Bowman S. R. et al. Generating Sentences from a Continuous Space, 2015

Decoder weakening

Free bits

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

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

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.

Kingma D. P. et al. Improving Variational Inference with Inverse Autoregressive Flow, 2016

Variational Lossy AutoEncoder, 2016

Lossy code via explicit information placement

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

- 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 $\mathbf{z}$.

Chen X. et al. Variational Lossy Autoencoder, 2016

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?

MNIST

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**

CIFAR10

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**

Chen X. et al. Variational Lossy Autoencoder, 2016

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

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

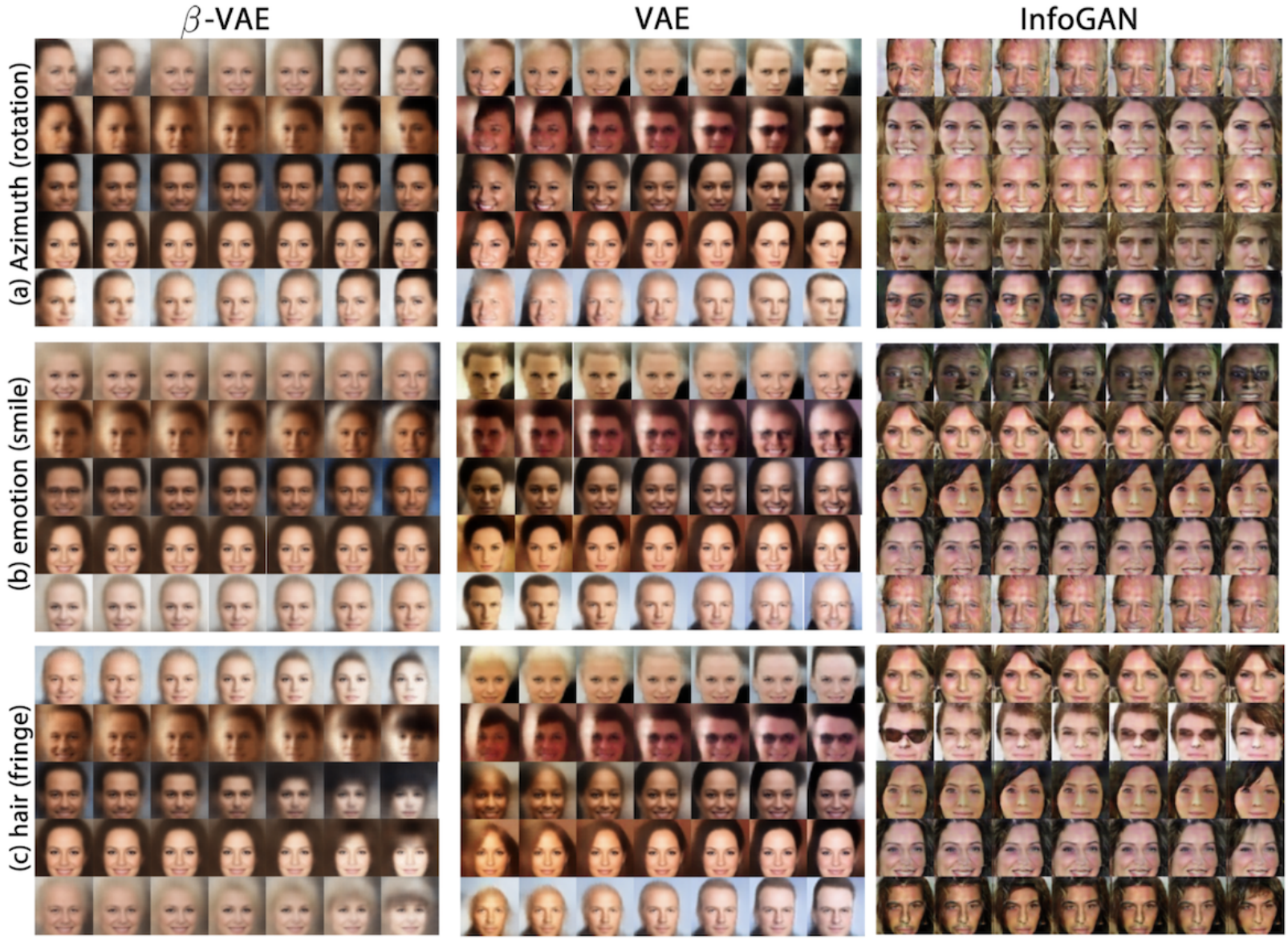
$$q(\mathbf{z}|\mathbf{x}) = \mathcal{N}\left(\mu(\mathbf{x})|\sigma^2(\mathbf{x})\right); \quad \mathbf{z}_{li} = \mu(\mathbf{x}_{li}); \quad \mathbf{z}_{lj} = \mu(\mathbf{x}_{lj}).$$

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

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

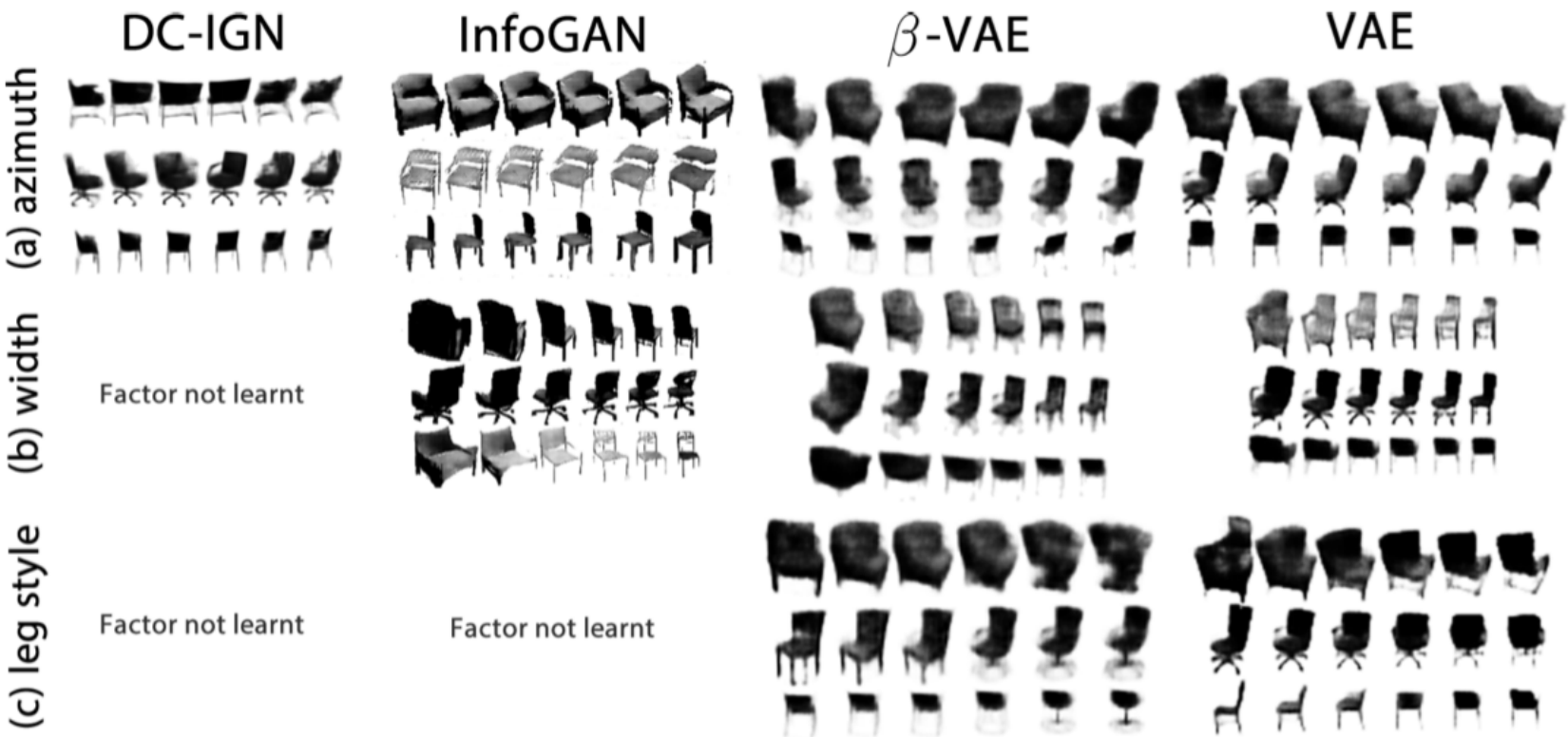
Higgins I. et al. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework, 2017

β-VAE



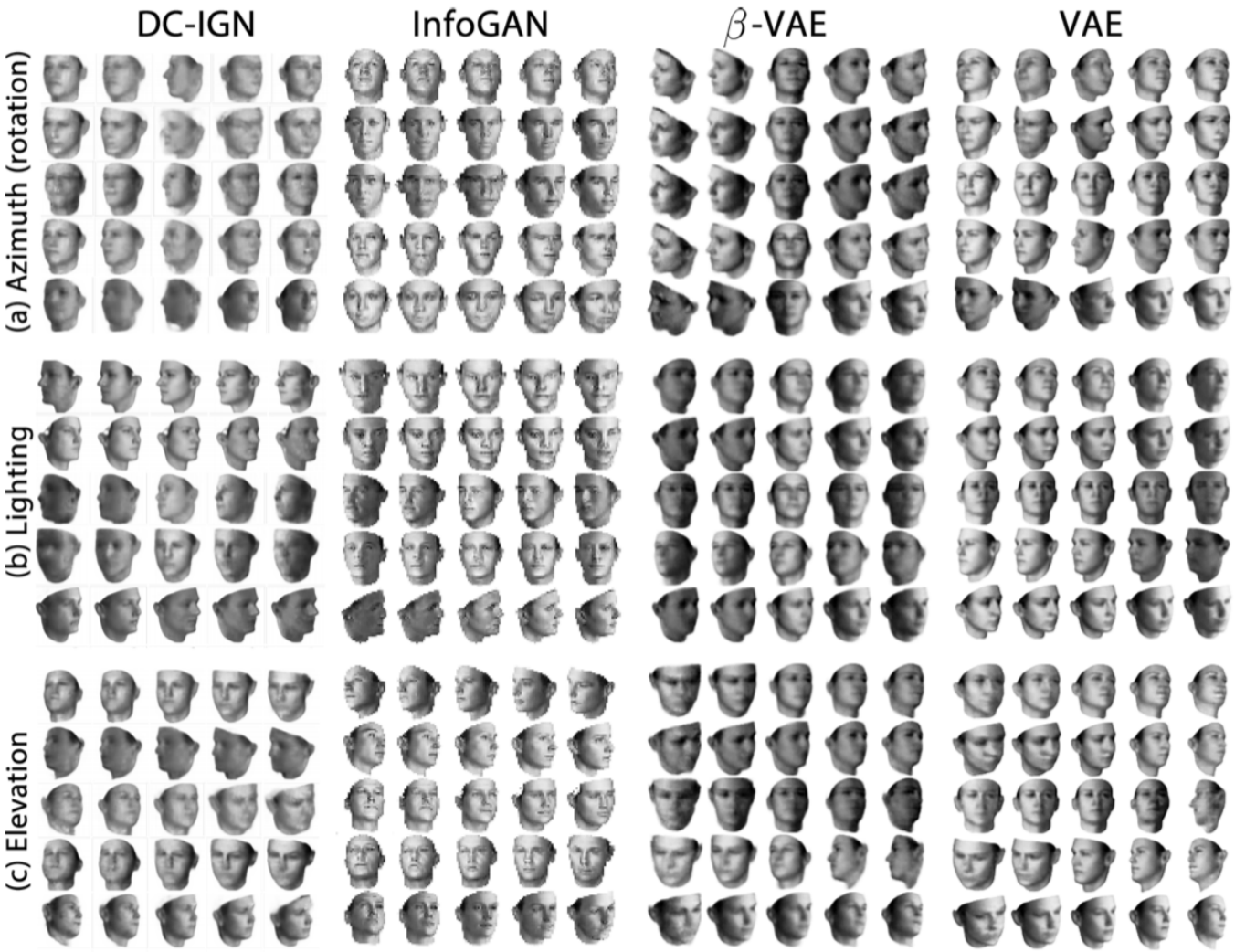
Higgins I. et al. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework, 2017

β-VAE



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

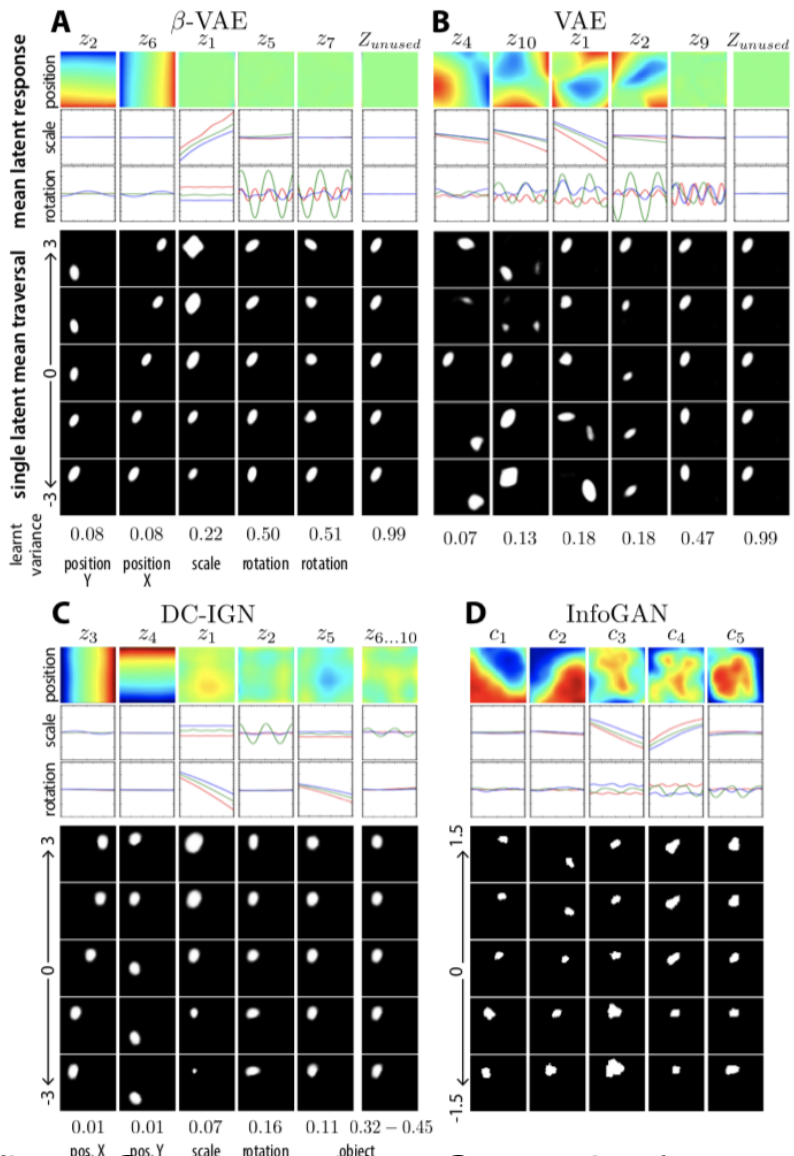
β-VAE



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	$\mathbf{99.23 \pm 0.1\%}$

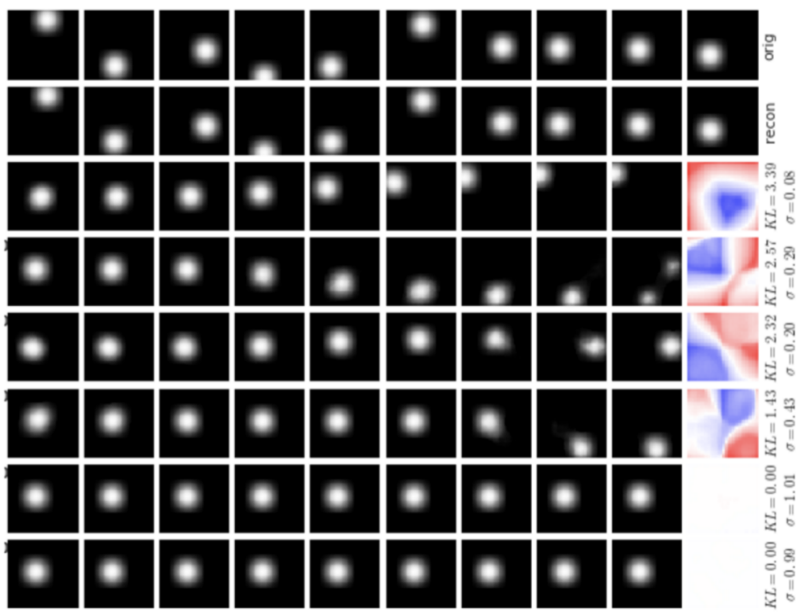


Higgins I. et al. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework, 2017

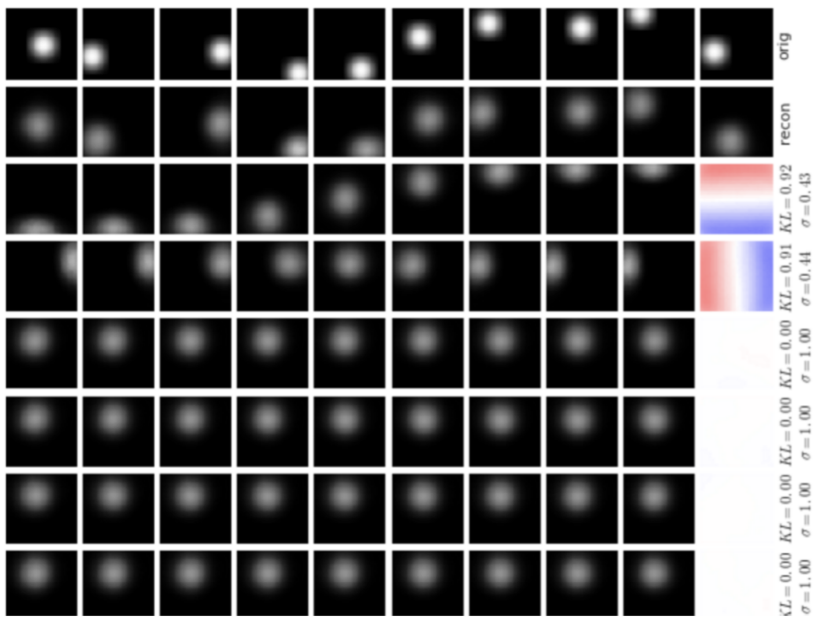
β-VAE

- **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

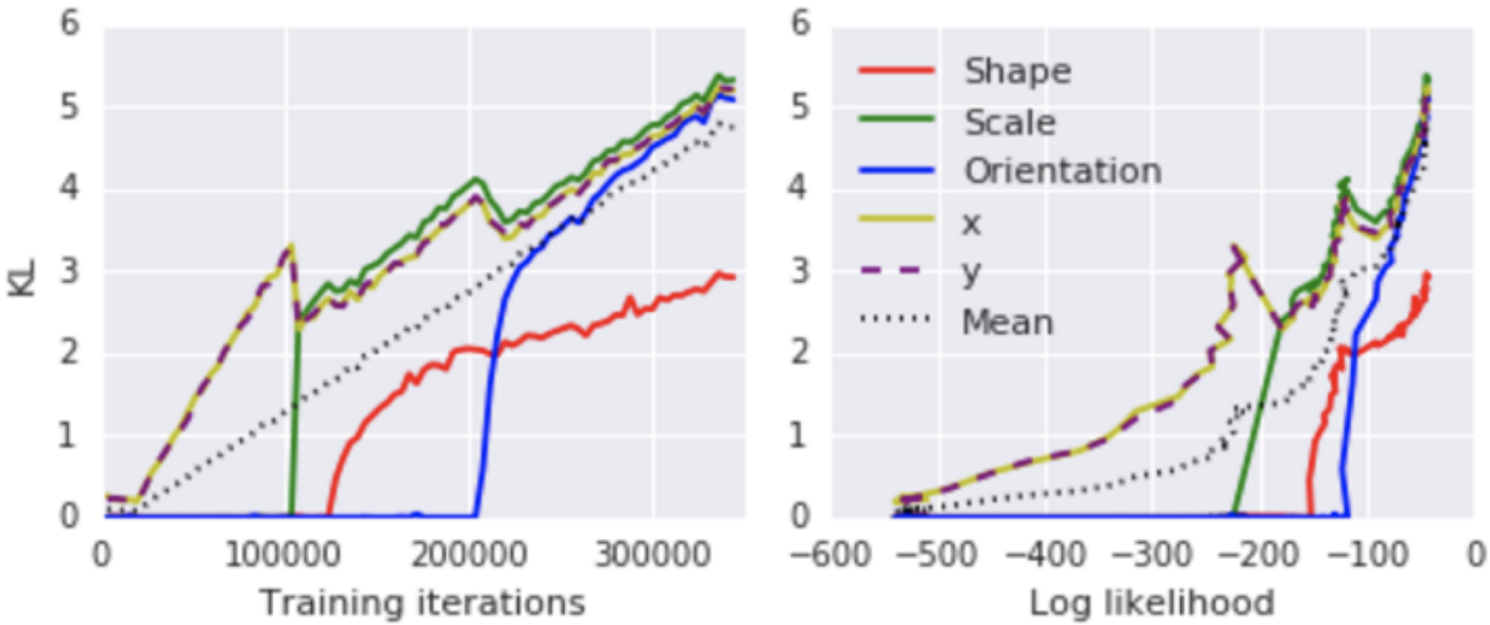


Burgess C. P. et al. Understanding disentangling in β-VAE, 2018

β-VAE

Controlled encoding capacity

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



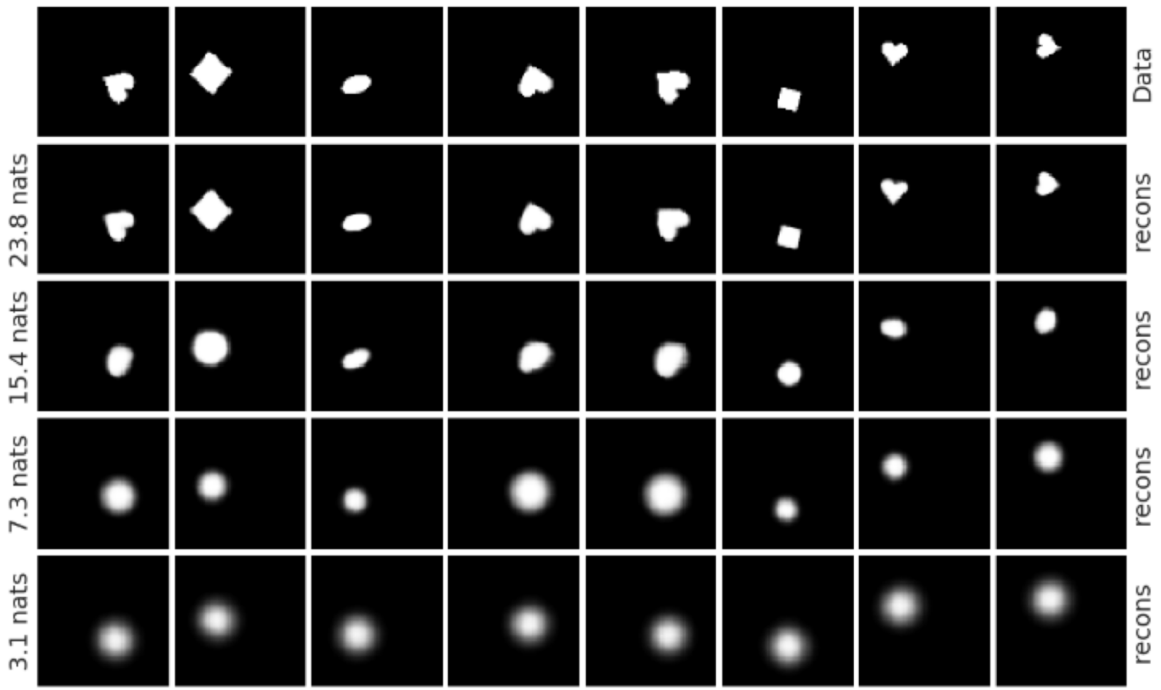
The early capacity is allocated to positional latents only, followed by a scale latent, then shape and orientation latents.

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

β-VAE

Controlled encoding capacity

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



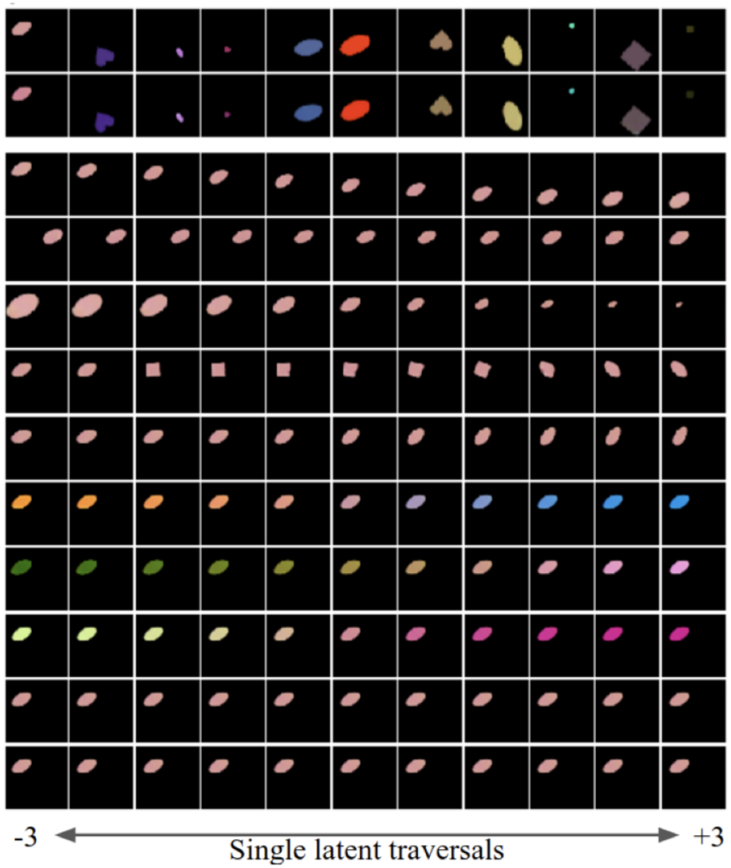
As the information capacity increases the different latents associated with their data generative factors become informative.

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

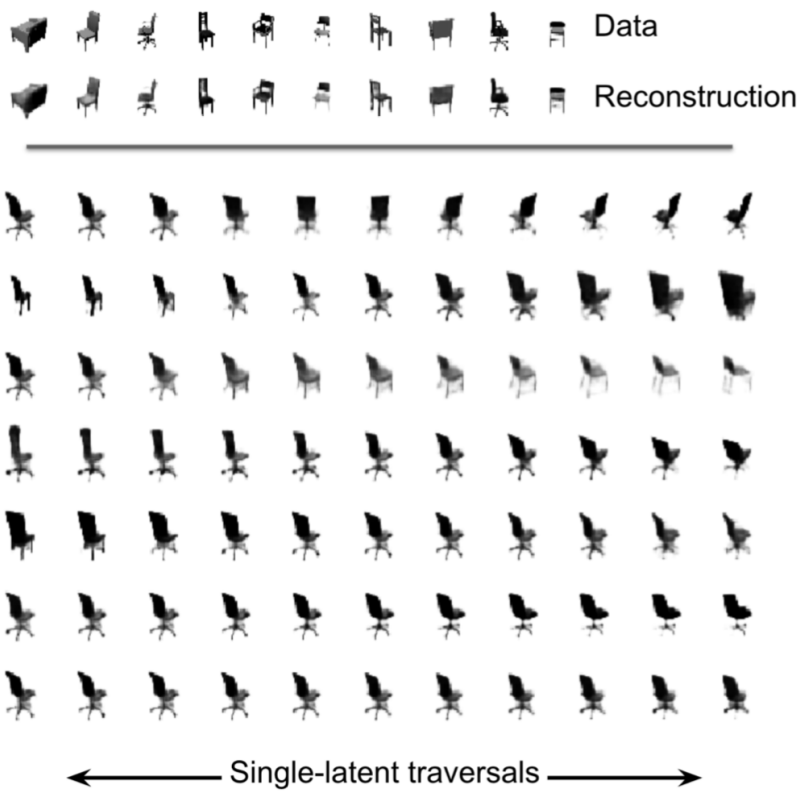
β-VAE

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

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

ELBO surgery

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

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

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

- $KL(q(\mathbf{z})||p(\mathbf{z}))$ 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

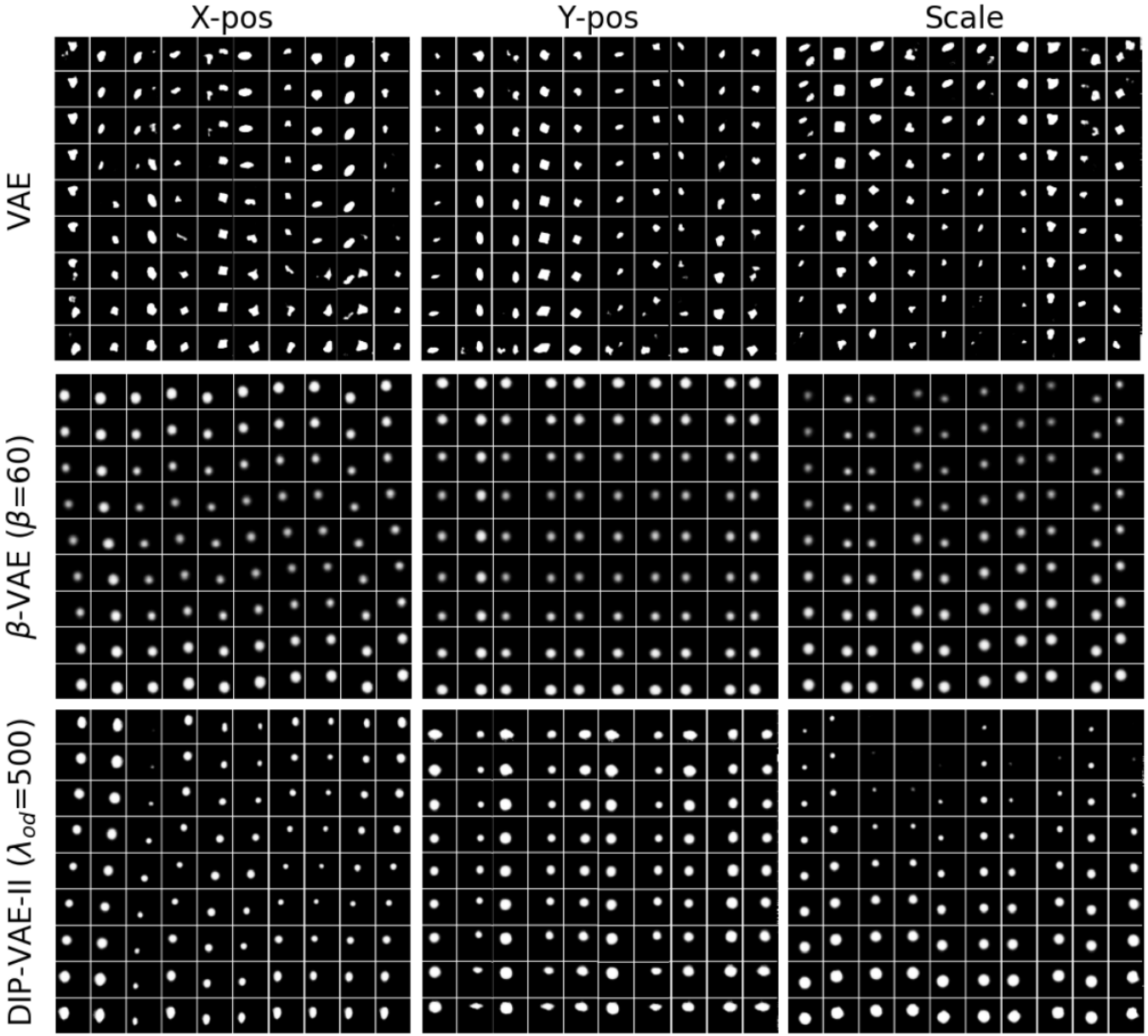
$$\max_{\theta, \phi} \text{ELBO}(\theta, \phi) - \lambda_1 \sum_{i \neq j} \left[\text{cov}_{\pi(\mathbf{x})} \boldsymbol{\mu}(\mathbf{x}) \right]_{ij}^2 - \lambda_2 \sum_i \left(\left[\text{cov}_{\pi(\mathbf{x})} \boldsymbol{\mu}(\mathbf{x}) \right]_{ii} - 1 \right)^2$$

DIP-VAE-II

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

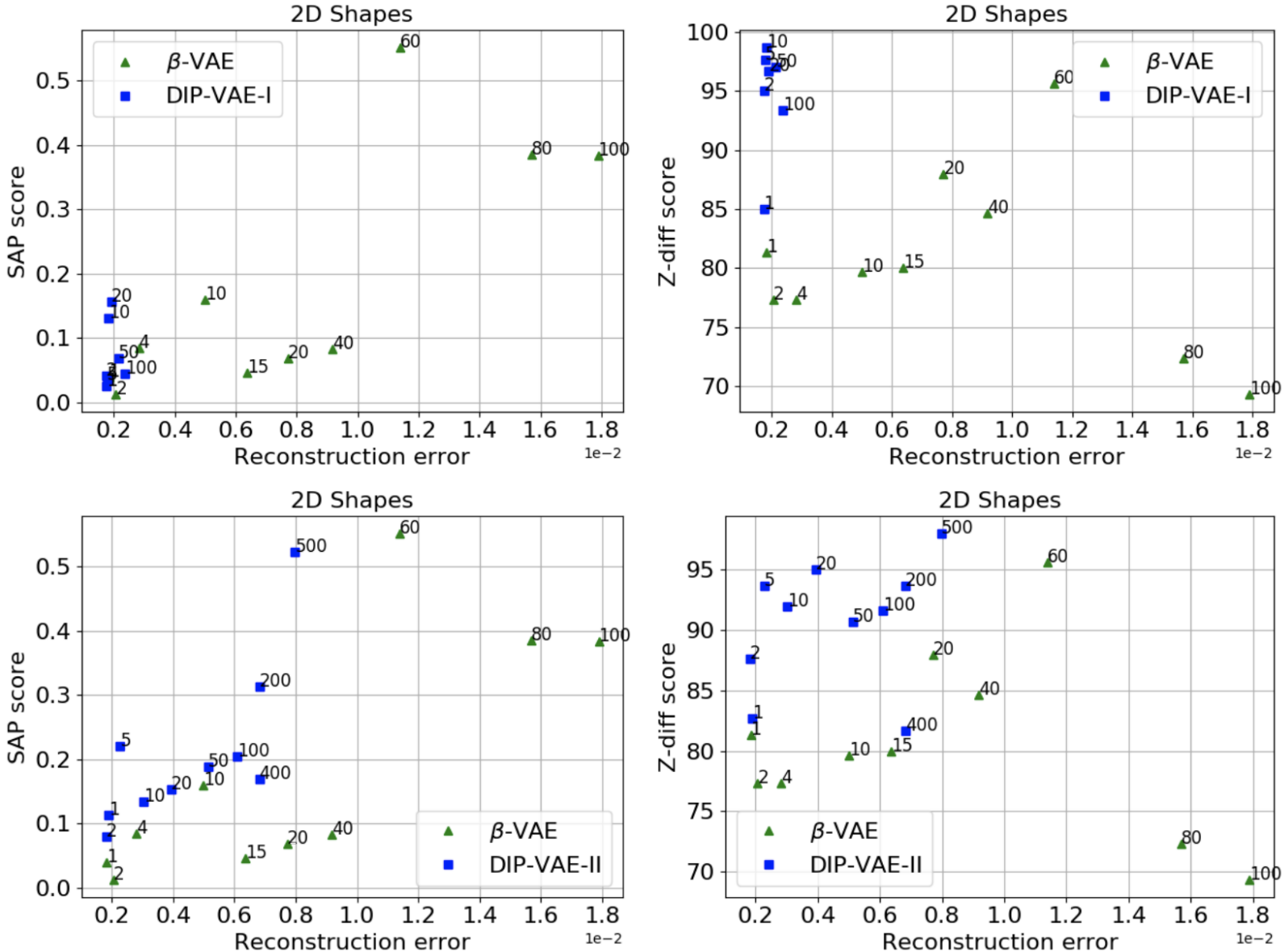
Kumar A., Sattigeri P., Balakrishnan A. Variational Inference of Disentangled Latent Concepts from Unlabeled Observations, 2017

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

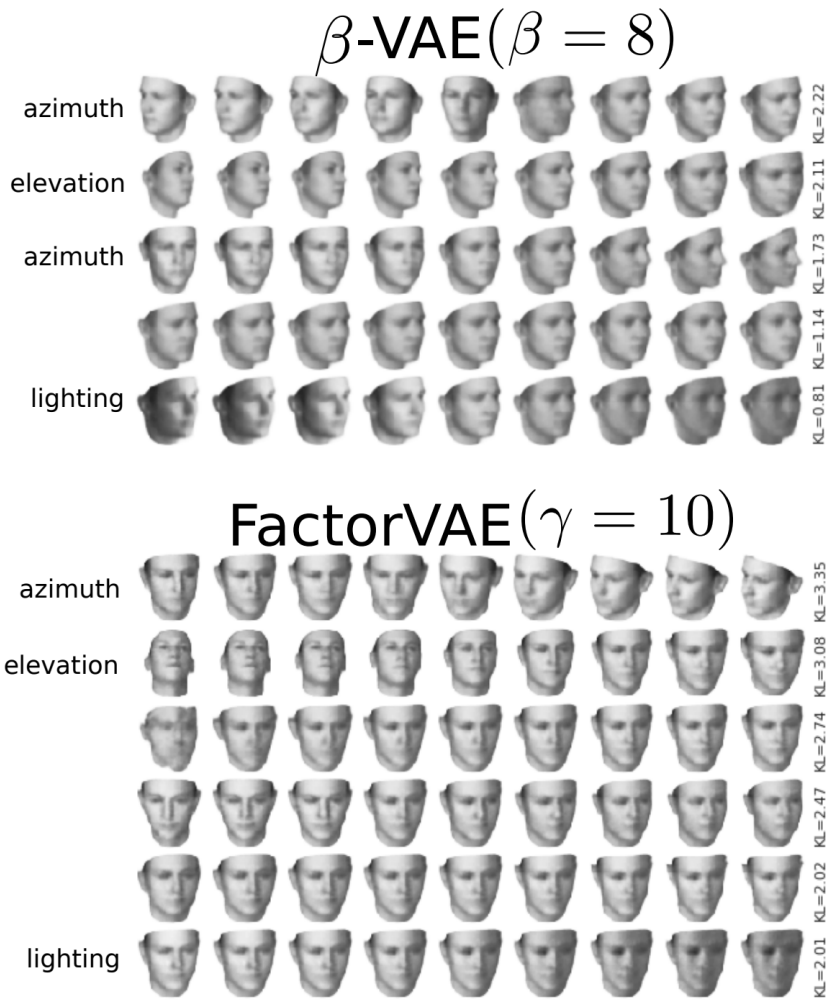
$$KL(q(\mathbf{z})||\prod_{j=1}^{d} q(z_j)) = KL(q(\mathbf{z})||\bar{q}(\mathbf{z})) =$$

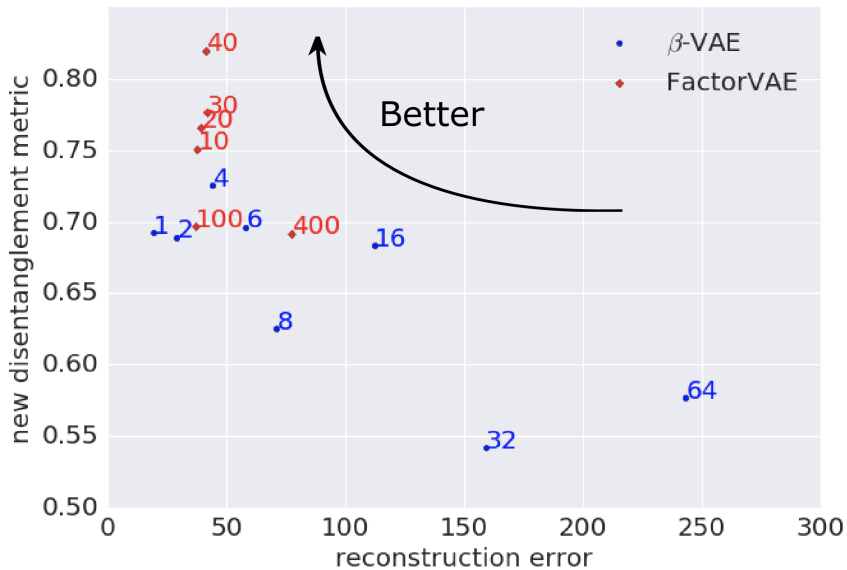
$$= \mathbb{E}_{q(\mathbf{z})} \log \frac{q(\mathbf{z})}{\bar{q}(\mathbf{z})} \approx \mathbb{E}_{q(\mathbf{z})} \log \frac{D(\mathbf{z})}{1 - D(\mathbf{z})}$$

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

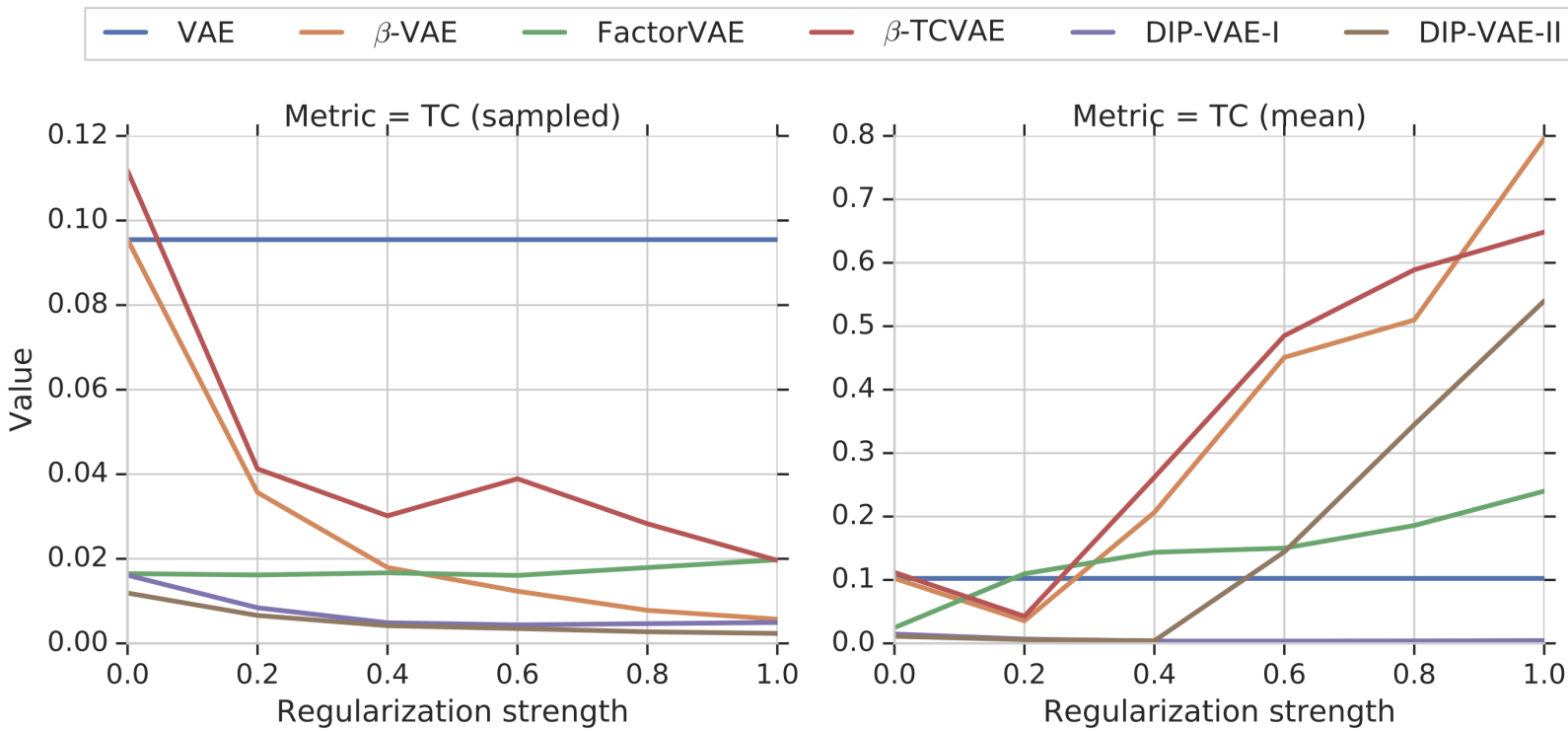
$$f(\mathbf{u}) = g^{-1}(h^{-1}(\mathbf{A}h(g(\mathbf{z})))).$$

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

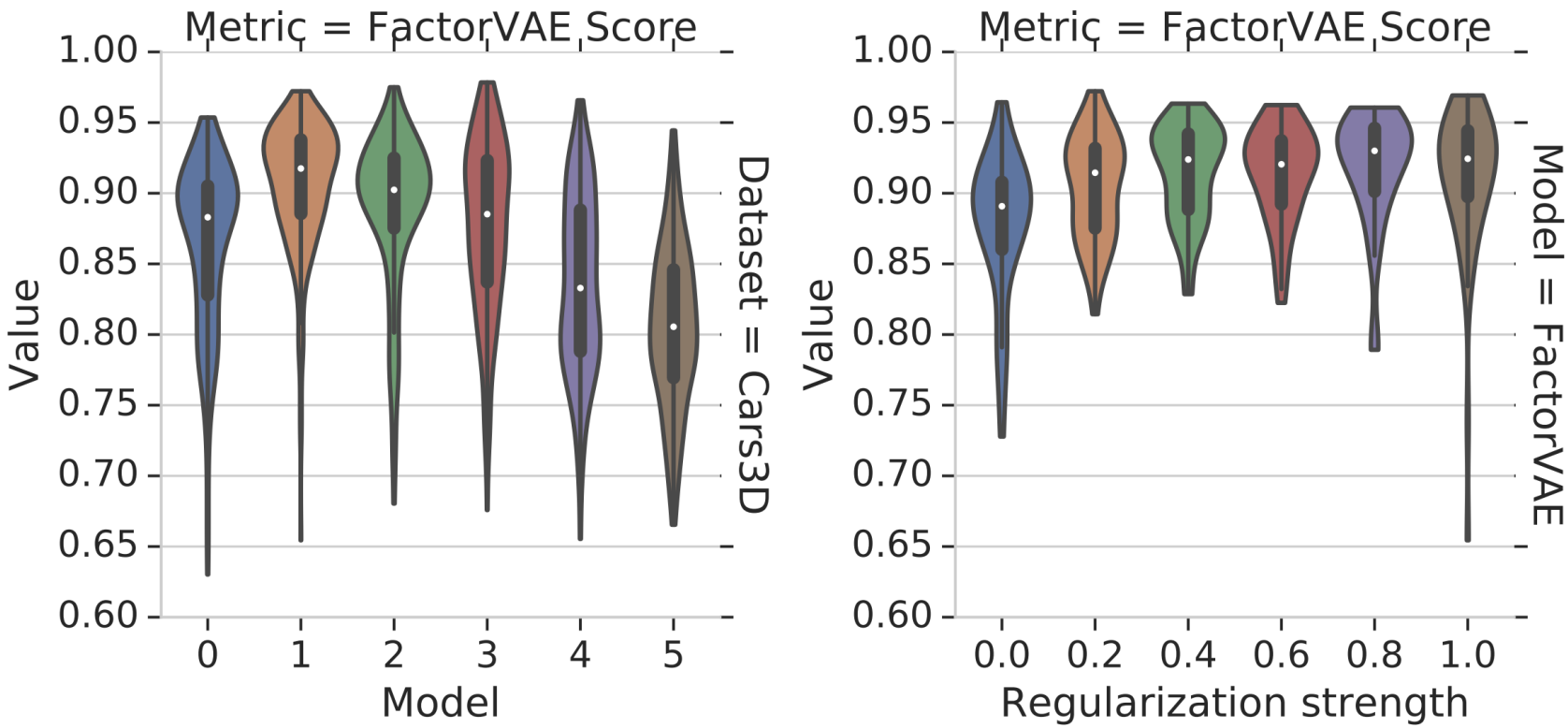
- **Training:** Factorizing **samples** from aggregated posterior $q(\mathbf{z}) = \prod_{i=1}^{d} q(z_i)$.
- **Inference:** Use a **mean** vector (usually mean of Gaussian encoder) as a representation.



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

Challenging Disentanglement Assumptions

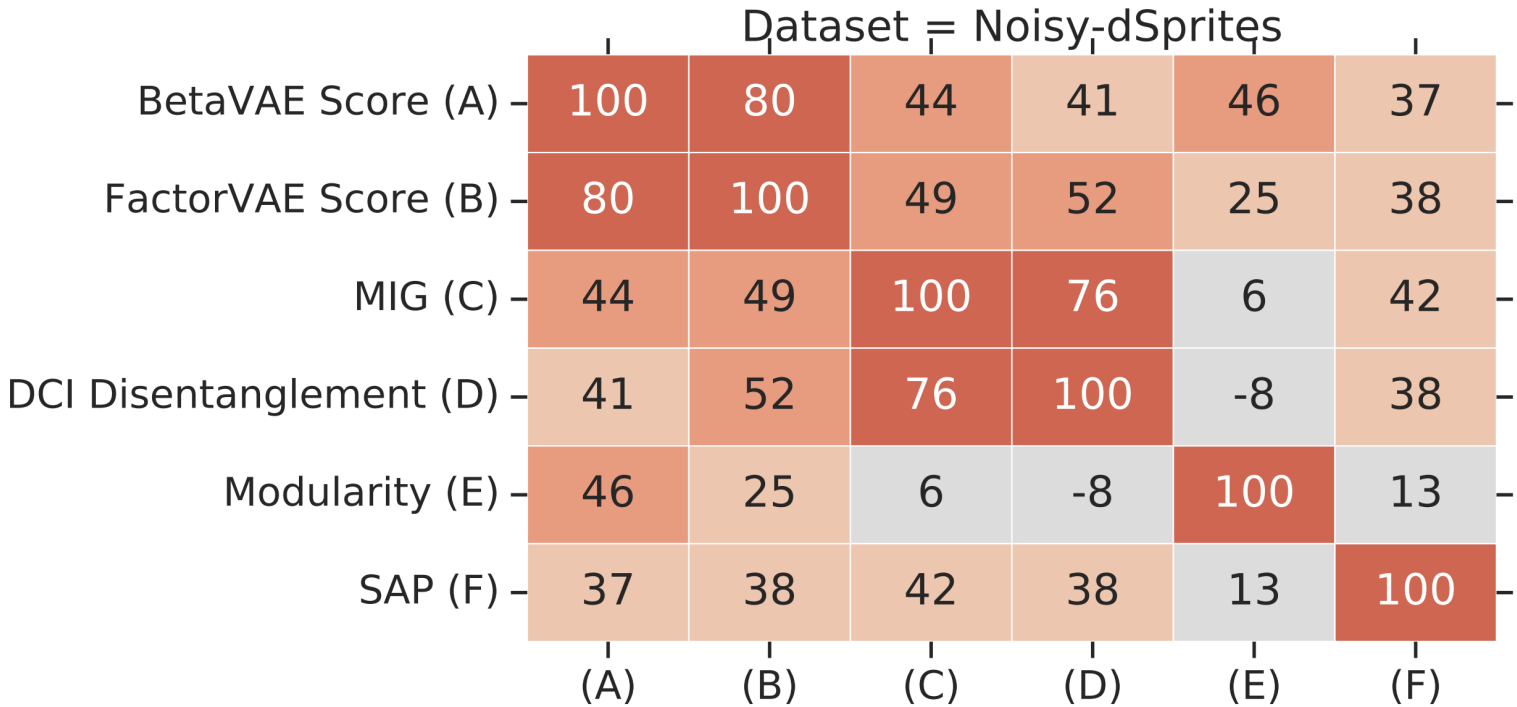
Importance of different models and hyperparameters for disentanglement



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

- VampPrior proposes to use a variational mixture of posteriors as the prior to approximate the aggregated posterior.

- The autoregressive flows could be used as the prior. This is equivalent to the use of the IAF posterior.
- More powerful decoder in VAE leads to more expressive generative model. However, too expressive decoder could lead to the posterior collapse.
- The decoder weakening is a set of techniques to avoid the posterior collapse.