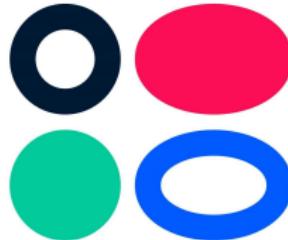


Deep Generative Models

Lecture 11

Roman Isachenko



Ozon Masters

Spring, 2021

VAE limitations

- ▶ Poor variational posterior distribution (encoder)

$$q(\mathbf{z}|\mathbf{x}, \phi) = \mathcal{N}(\mathbf{z}|\mu_\phi(\mathbf{x}), \sigma_\phi^2(\mathbf{x})).$$

- ▶ Poor prior distribution

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

- ▶ Poor probabilistic model (decoder)

$$p(\mathbf{x}|\mathbf{z}, \theta) = \mathcal{N}(\mathbf{x}|\mu_\theta(\mathbf{z}), \sigma_\theta^2(\mathbf{z})).$$

- ▶ Loose lower bound

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

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

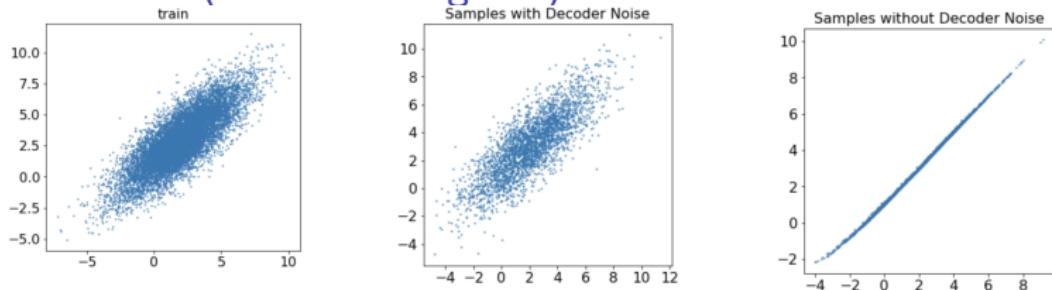
Let data distribution is $\pi(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Possible cases:

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

What is the difference?

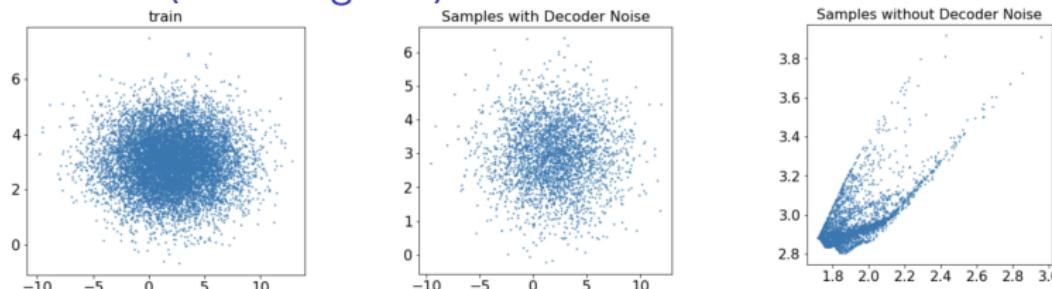
Posterior collapse: toy example

Multivariate (Σ is non-diagonal)



The encoder uses latent variables to model data.

Univariate (Σ 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".

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

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

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

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

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

- ▶ More powerful $p(\mathbf{x}|\mathbf{z}, \theta)$ leads to more powerful generative model $p(\mathbf{x}|\theta)$.
- ▶ Too powerful $p(\mathbf{x}|\mathbf{z}, \theta)$ 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 $p(\mathbf{x}|\mathbf{z}, \theta)$ more powerful?

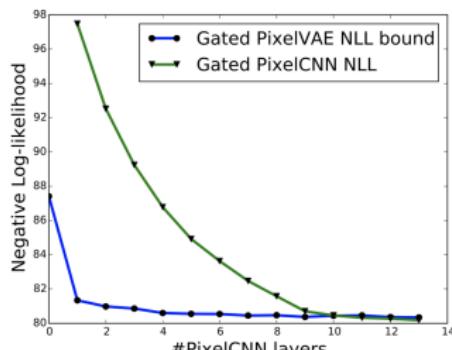
Autoregressive decoder

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

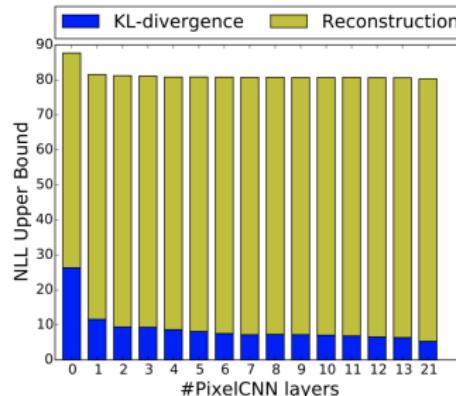
PixelVAE

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)

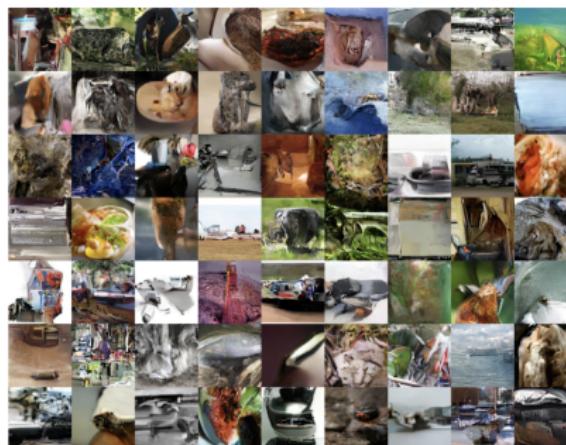
PixelVAE

MNIST

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

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) | $= \mathbf{3.57} (3.48)$ |
| Hierarchical PixelVAE | $\leq 3.66 (3.59)$ |



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.

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 .

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

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.

Variational Lossy AutoEncoder, 2016

Lossy code via explicit information placement

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

- ▶ $\text{WindowAround}(i)$ 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} .

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?

CIFAR10

MNIST

| Model | NLL Test |
|---|--------------|
| Normalizing flows (Rezende & Mohamed, 2015) | 85.10 |
| DRAW (Gregor et al., 2015) | < 80.97 |
| Discrete VAE (Rollef, 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 |
|---|-----------------|
| <i>Results with tractable likelihood models:</i> | |
| 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 |
| <i>Results with variationally trained latent-variable models:</i> | |
| 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 |

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.

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} .

β -VAE

Constrained optimization

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

Objective

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

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.

β -VAE

Disentangling metric

1. Generate two sets of objects

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

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

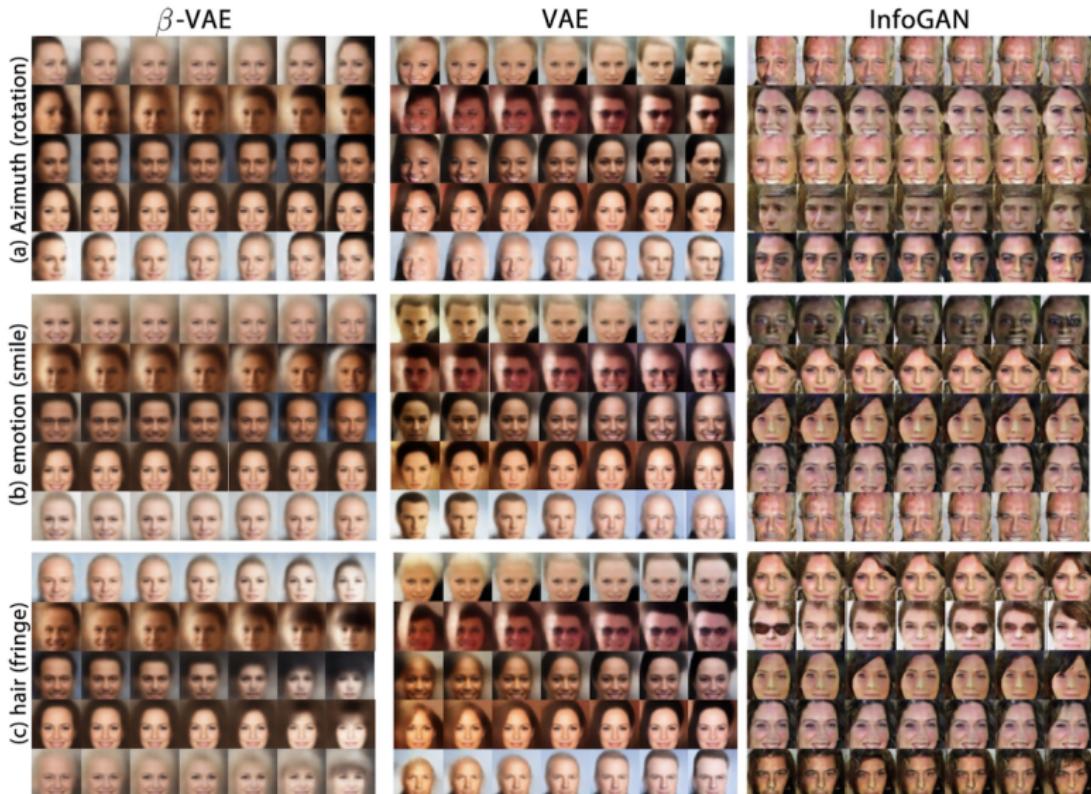
2. Find representations

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

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

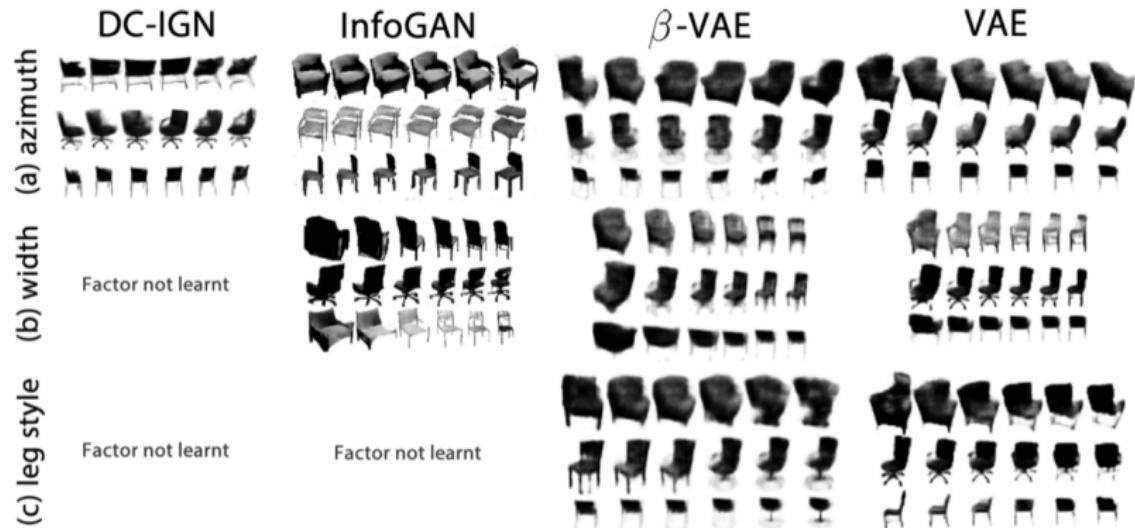
$$\mathbf{z}_{\text{diff}} = \frac{1}{L} \sum_{l=1}^L |\mathbf{z}_{li} - \mathbf{z}_{lj}|.$$

β -VAE

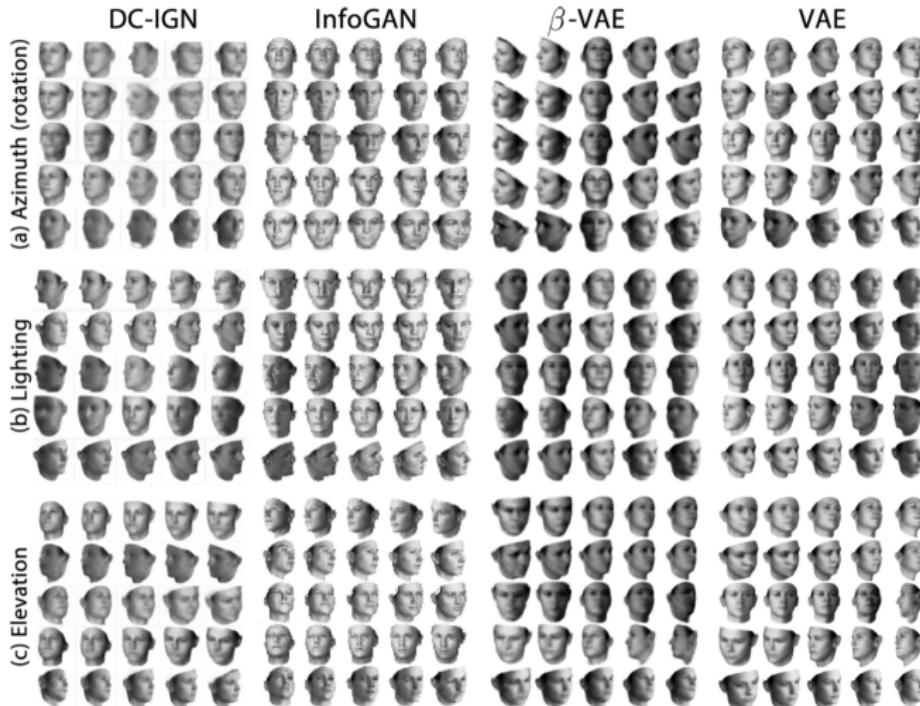


Higgins I. et al. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework, 2017

β -VAE

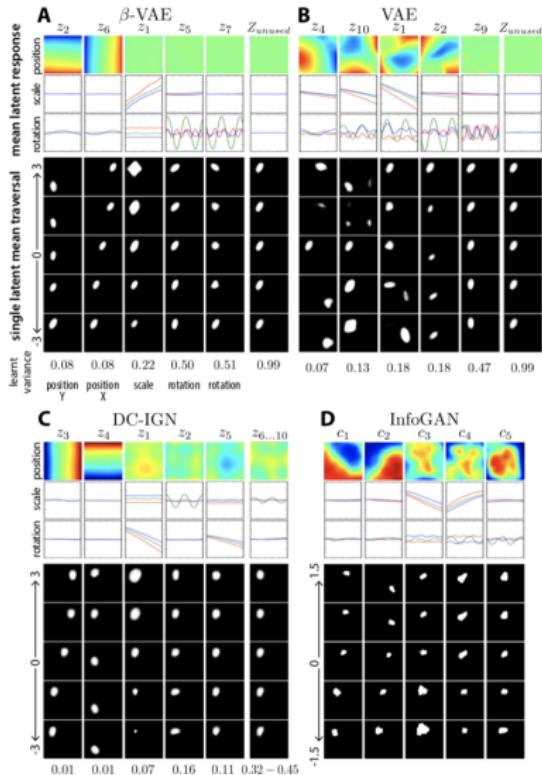


β -VAE



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 | $99.3 \pm 0.1\%$ |
| InfoGAN | $73.5 \pm 0.9\%$ |
| VAE untrained | $44.14 \pm 2.5\%$ |
| VAE | $61.58 \pm 0.5\%$ |
| β -VAE | $99.23 \pm 0.1\%$ |

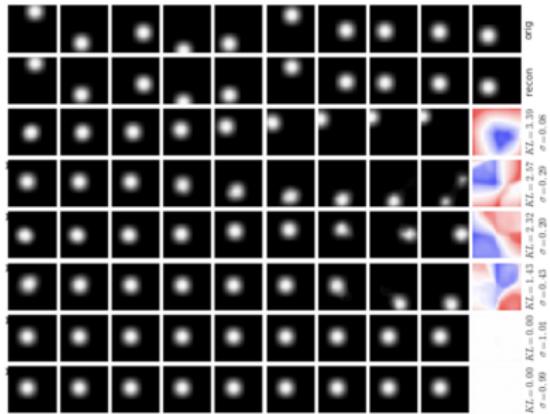


Higgins I. et al. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework, 2017

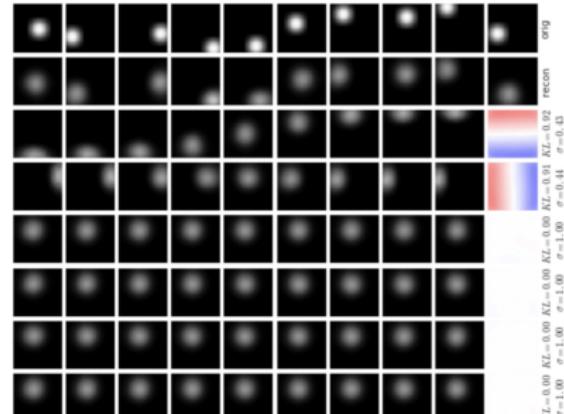
β -VAE

- ▶ **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.

$\beta = 1$



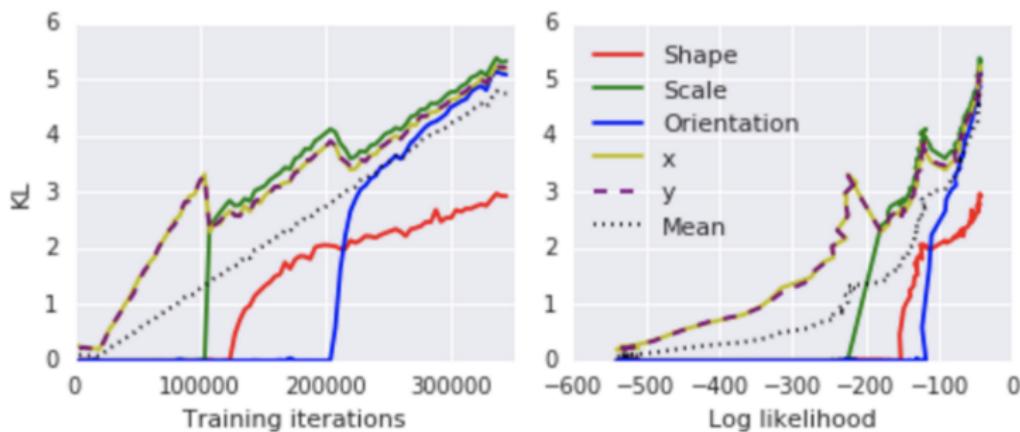
$\beta = 150$



β -VAE

Controlled encoding capacity

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

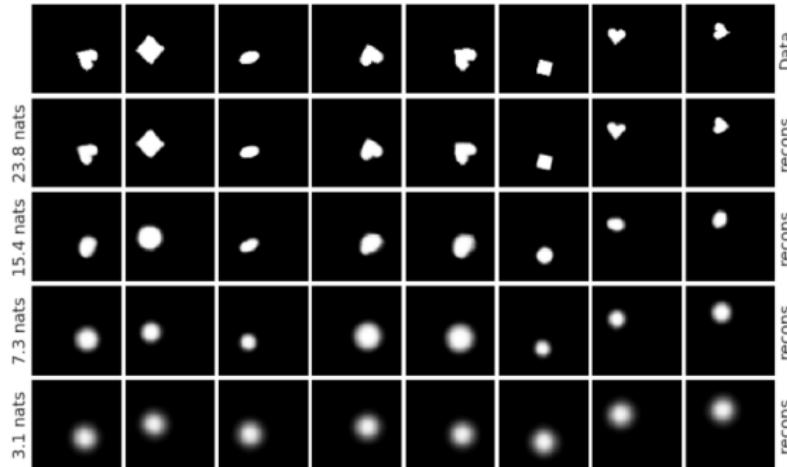


The early capacity is allocated to positional latents only, followed by a scale latent, then shape and orientation latents.

β -VAE

Controlled encoding capacity

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

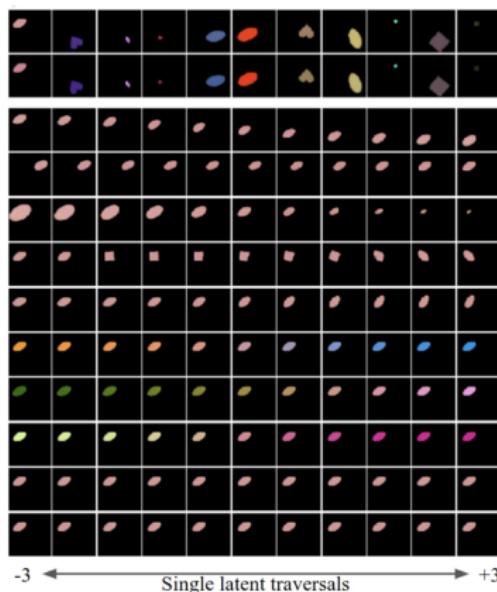


As the information capacity increases the different latents associated with their data generative factors become informative.

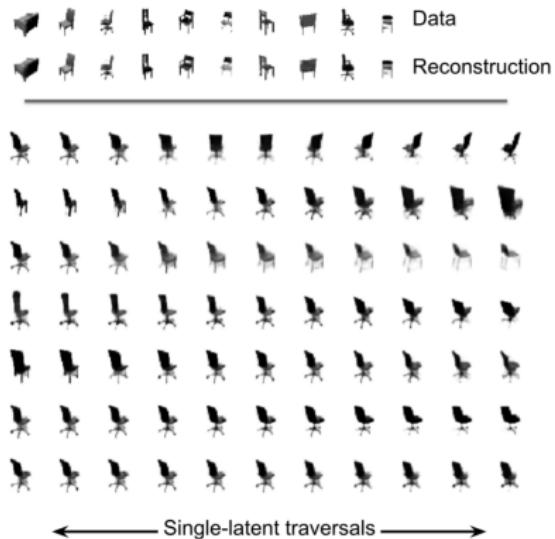
β -VAE

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

(a) Coloured dSprites



(b) 3D Chairs



β -VAE

ELBO

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

ELBO surgery

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

Minimization of MI

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

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

$$\begin{aligned}\mathcal{L}(q, \theta) &= \underbrace{\mathbb{E}_{\pi(\mathbf{x})} [\mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \log p(\mathbf{x}|\mathbf{z}, \theta) - KL(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))]}_{\text{ELBO}} - \lambda \cdot KL(q(\mathbf{z})||p(\mathbf{z})) = \\ &= \underbrace{\mathbb{E}_{\pi(\mathbf{x})} [\mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \log p(\mathbf{x}|\mathbf{z}, \theta)]}_{\text{Reconstruction loss}} - \underbrace{\mathbb{I}_{q(i, \mathbf{z})}[i, \mathbf{z}]}_{\text{Mutual info}} - (1 + \lambda) \cdot \underbrace{KL(q(\mathbf{z})||p(\mathbf{z}))}_{\text{Marginal KL}}\end{aligned}$$

DIP-VAE

DIP-VAE Objective

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

- ▶ $KL(q(z)||p(z))$ is intractable.
- ▶ Let match the moments of $q(z)$ and $p(z)$.

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

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

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

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

DIP-VAE

DIP-VAE Objective

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

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

DIP-VAE-I

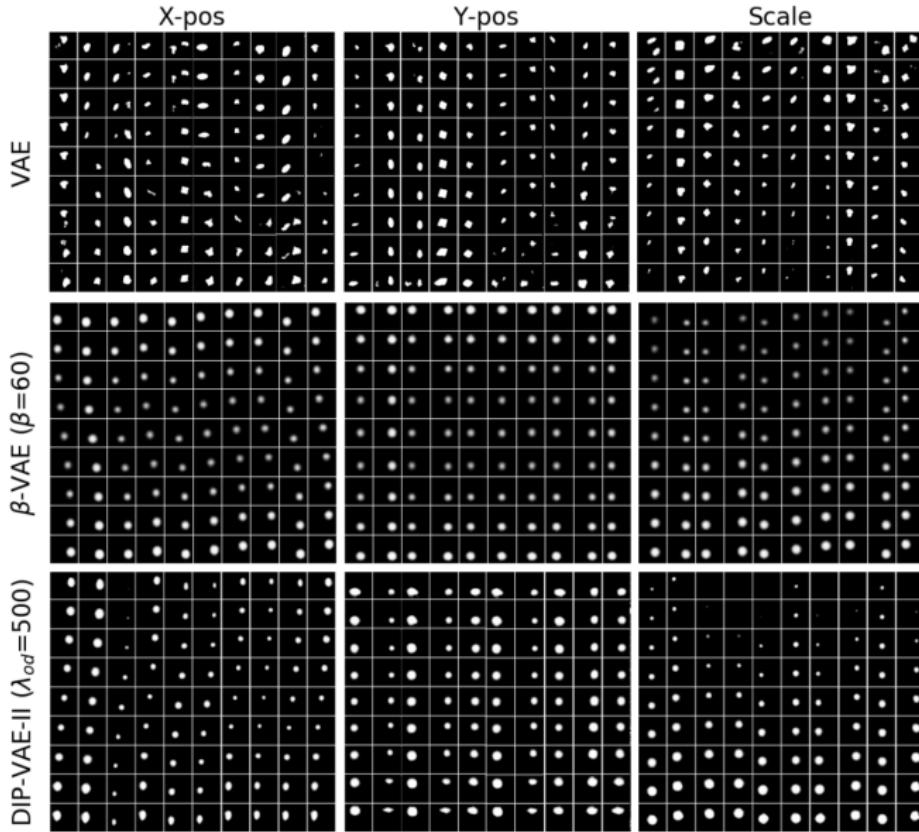
$$\max_{\theta, \phi} \text{ELBO}(\theta, \phi) - \lambda_1 \sum_{i \neq j} [\text{cov}_{\pi(x)} \mu(x)]_{ij}^2 - \lambda_2 \sum_i ([\text{cov}_{\pi(x)} \mu(x)]_{ii} - 1)^2$$

DIP-VAE-II

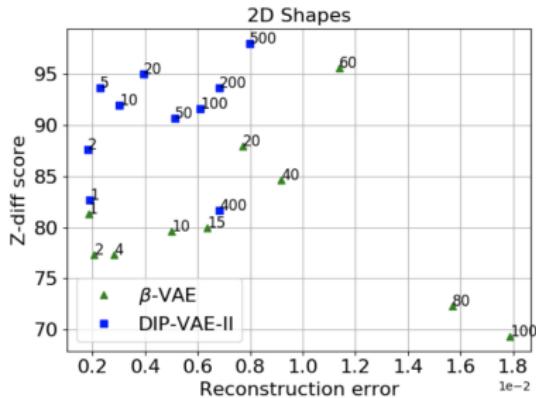
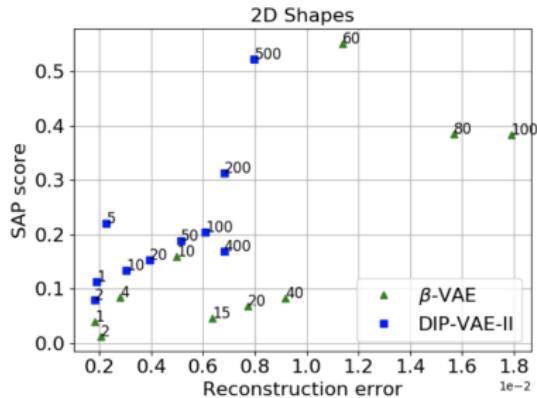
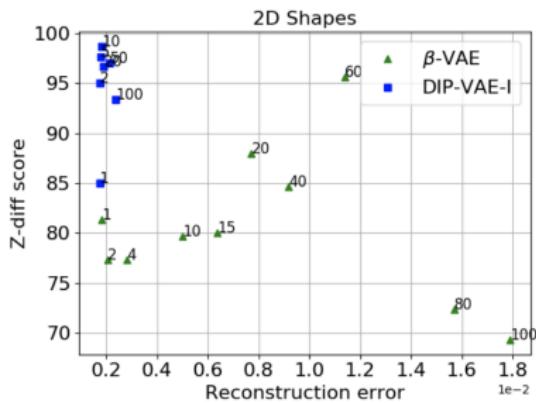
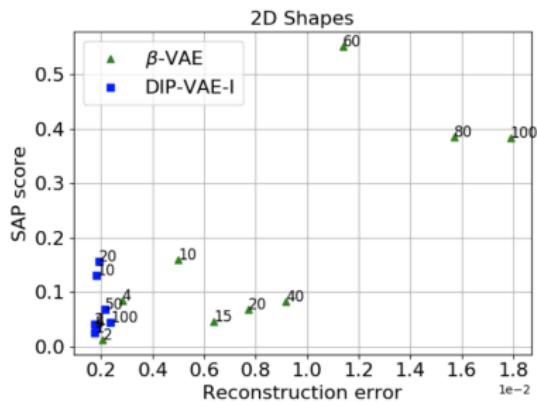
$$\max_{\theta, \phi} \text{ELBO}(\theta, \phi) - \lambda_1 \sum_{i \neq j} [\text{cov}_{q(z)}(z)]_{ij}^2 - \lambda_2 \sum_i ([\text{cov}_{q(z)}(z)]_{ii} - 1)^2$$

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

DIP-VAE



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

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

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

FactorVAE

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

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

Density ratio trick

$$\begin{aligned} \frac{q_1(\mathbf{x})}{q_2(\mathbf{x})} &= \frac{p(\mathbf{x}|y=1)}{p(\mathbf{x}|y=0)} = \frac{p(y=1|\mathbf{x})p(\mathbf{x})}{p(y=1)} \Big/ \frac{p(y=0|\mathbf{x})p(\mathbf{x})}{p(y=0)} = \\ &= \frac{p(y=1|\mathbf{x})}{p(y=0|\mathbf{x})} = \frac{p(y=1|\mathbf{x})}{1 - p(y=1|\mathbf{x})} = \frac{D(\mathbf{x})}{1 - D(\mathbf{x})} \end{aligned}$$

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 $q_2(\mathbf{x})$.

FactorVAE

FactorVAE objective

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

Total correlation regularizer

$$\begin{aligned} KL(q(\mathbf{z}) || \prod_{j=1}^d q(z_j)) &= KL(q(\mathbf{z}) || \bar{q}(\mathbf{z})) = \\ &= \mathbb{E}_{q(\mathbf{z})} \log \frac{q(\mathbf{z})}{\bar{q}(\mathbf{z})} \approx \mathbb{E}_{q(\mathbf{z})} \log \frac{D(\mathbf{z})}{1 - D(\mathbf{z})} \end{aligned}$$

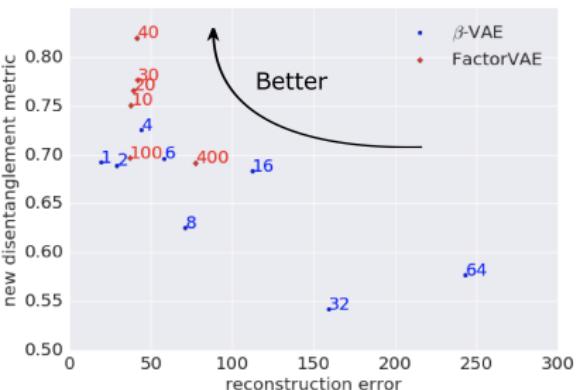
VAE and GAN are trained simultaneously.

FactorVAE

β -VAE ($\beta = 8$)



FactorVAE ($\gamma = 10$)



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

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

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

Challenging Disentanglement Assumptions

Proof (1)

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

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

- ▶ g is bijective (since $p(\mathbf{z}) = \prod_{i=1}^d dp(z_i)$).
 - ▶ $\frac{\partial g_i(\mathbf{u})}{\partial u_i} \neq 0$, 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 $h : (0, 1]^d \rightarrow \mathbb{R}^d$

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

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.

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

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

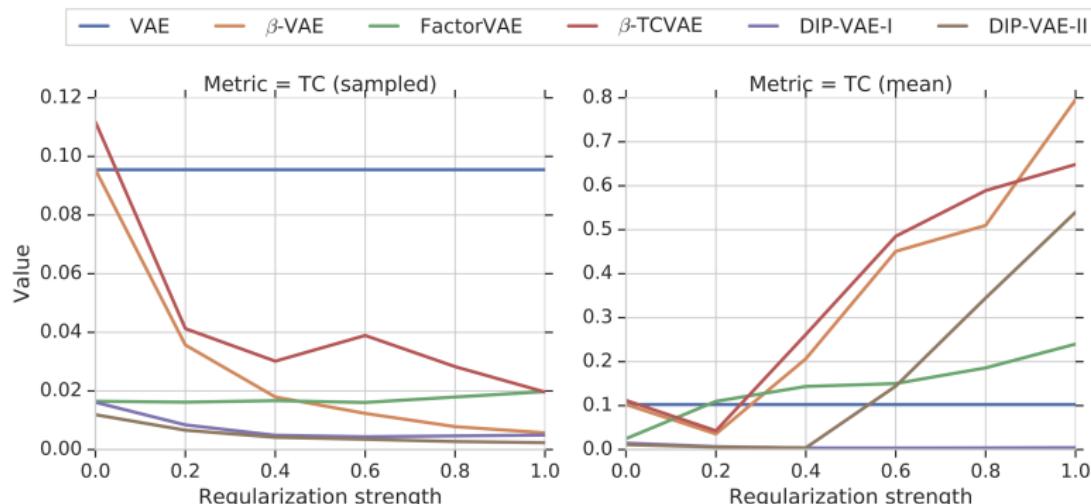
By definition $f(\mathbf{z})$ has the same marginal distribution as \mathbf{z} :

$$P(\mathbf{z} \leq \mathbf{u}) = P(f(\mathbf{z}) \leq \mathbf{u}) \text{ and } \frac{\partial f_i(\mathbf{u})}{\partial u_j} \neq 0.$$

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

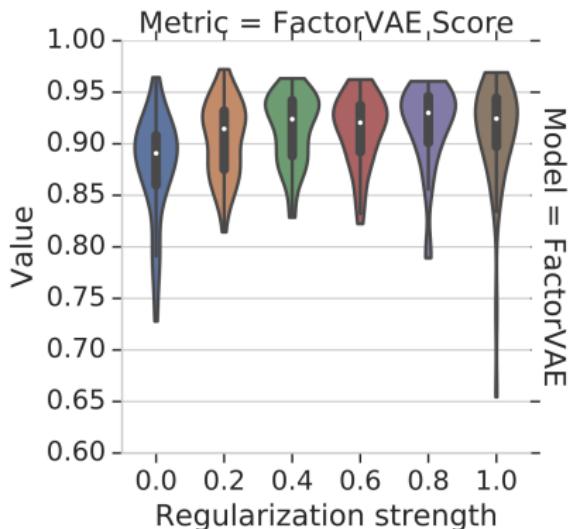
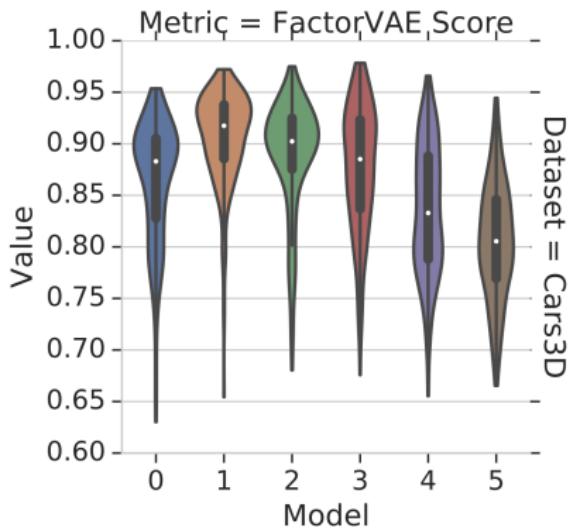
Challenging Disentanglement Assumptions

- ▶ **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.



Challenging Disentanglement Assumptions

Importance of different models and hyperparameters for disentanglement



Challenging Disentanglement Assumptions

Agreement of different disentanglement metrics

| | Dataset = Noisy-dSprites | | | | | |
|-------------------------|--------------------------|-----|-----|-----|-----|-----|
| | (A) | (B) | (C) | (D) | (E) | (F) |
| BetaVAE Score (A) | 100 | 80 | 44 | 41 | 46 | 37 |
| FactorVAE Score (B) | 80 | 100 | 49 | 52 | 25 | 38 |
| MIG (C) | 44 | 49 | 100 | 76 | 6 | 42 |
| DCI Disentanglement (D) | 41 | 52 | 76 | 100 | -8 | 38 |
| Modularity (E) | 46 | 25 | 6 | -8 | 100 | 13 |
| SAP (F) | 37 | 38 | 42 | 38 | 13 | 100 |

Summary

- ▶ 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.