

# Deep Generative Models

Roman Isachenko

Moscow Institute of Physics and Technology

2019

## VAE limitations

- ▶ Poor variational posterior distribution (encoder)

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

- ▶ Poor prior distribution

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

- ▶ Poor probabilistic model (decoder)

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

- ▶ Loose lower bound

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

# Importance Sampling

$$\begin{aligned} p(\mathbf{x}|\theta) &= \int p(\mathbf{x}, \mathbf{z}|\theta) d\mathbf{z} = \int \left[ \frac{p(\mathbf{x}, \mathbf{z}|\theta)}{q(\mathbf{z}|\mathbf{x})} \right] q(\mathbf{z}|\mathbf{x}) d\mathbf{z} \\ &= \int f(\mathbf{x}, \mathbf{z}) q(\mathbf{z}|\mathbf{x}) d\mathbf{z} = \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z}|\mathbf{x})} f(\mathbf{x}, \mathbf{z}) \end{aligned}$$

Here  $f(\mathbf{x}, \mathbf{z}) = \frac{p(\mathbf{x}, \mathbf{z}|\theta)}{q(\mathbf{z}|\mathbf{x})}$ .

$$\begin{aligned} \log p(\mathbf{x}|\theta) &= \log \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z}|\mathbf{x})} f(\mathbf{x}, \mathbf{z}) \geq \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z}|\mathbf{x})} \log f(\mathbf{x}, \mathbf{z}) = \\ &= \mathbb{E}_{\mathbf{z} \sim q(\mathbf{z}|\mathbf{x})} \log \frac{p(\mathbf{x}, \mathbf{z}|\theta)}{q(\mathbf{z}|\mathbf{x})} = \mathcal{L}(q, \theta). \end{aligned}$$

Could we choose better  $f(\mathbf{x}, \mathbf{z})$ ?

Let  $f(\mathbf{x}, \mathbf{z}_1, \dots, \mathbf{z}_K) = \frac{1}{K} \sum_{k=1}^K \frac{p(\mathbf{x}, \mathbf{z}_k|\theta)}{q(\mathbf{z}_k|\mathbf{x})}$ .

$$f(\mathbf{x}, \mathbf{z}_1, \dots, \mathbf{z}_K) = \frac{1}{K} \sum_{k=1}^K \frac{p(\mathbf{x}, \mathbf{z}_k | \boldsymbol{\theta})}{q(\mathbf{z}_k | \mathbf{x})}$$

$$\begin{aligned} \log p(\mathbf{x} | \boldsymbol{\theta}) &= \log \mathbb{E}_{\mathbf{z}_1, \dots, \mathbf{z}_K \sim q(\mathbf{z} | \mathbf{x})} f(\mathbf{x}, \mathbf{z}, \dots, \mathbf{z}_K) \geq \\ &\geq \mathbb{E}_{\mathbf{z}_1, \dots, \mathbf{z}_K \sim q(\mathbf{z} | \mathbf{x})} \log f(\mathbf{x}, \mathbf{z}, \dots, \mathbf{z}_K) = \\ &= \mathbb{E}_{\mathbf{z}_1, \dots, \mathbf{z}_K \sim q(\mathbf{z} | \mathbf{x})} \log \left[ \frac{1}{K} \sum_{k=1}^K \frac{p(\mathbf{x}, \mathbf{z}_k | \boldsymbol{\theta})}{q(\mathbf{z}_k | \mathbf{x})} \right] = \mathcal{L}_K(q, \boldsymbol{\theta}). \end{aligned}$$

---

<https://arxiv.org/pdf/1509.00519.pdf>

## VAE objective

$$p(\mathbf{x}|\boldsymbol{\theta}) \geq \mathcal{L}(q, \boldsymbol{\theta}) = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \log \frac{p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z}|\mathbf{x})} \rightarrow \max_{\phi, \boldsymbol{\theta}}$$

$$\mathcal{L}(q, \boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z}_1, \dots, \mathbf{z}_K \sim q(\mathbf{z}|\mathbf{x})} \left( \frac{1}{K} \sum_{k=1}^K \log \frac{p(\mathbf{x}, \mathbf{z}_k|\boldsymbol{\theta})}{q(\mathbf{z}_k|\mathbf{x})} \right) \rightarrow \max_{q, \boldsymbol{\theta}}.$$

## IWAE objective

$$\mathcal{L}_K(q, \boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z}_1, \dots, \mathbf{z}_K \sim q(\mathbf{z}|\mathbf{x})} \log \left( \frac{1}{K} \sum_{k=1}^K \frac{p(\mathbf{x}, \mathbf{z}_k|\boldsymbol{\theta})}{q(\mathbf{z}_k|\mathbf{x})} \right) \rightarrow \max_{q, \boldsymbol{\theta}}.$$

If  $K = 1$ , these objectives coincide.

## Theorem

1.  $\log p(\mathbf{x}|\boldsymbol{\theta}) \geq \mathcal{L}_K(q, \boldsymbol{\theta}) \geq \mathcal{L}_M(q, \boldsymbol{\theta})$ , for  $K \geq M$ ;
2.  $\log p(\mathbf{x}|\boldsymbol{\theta}) = \lim_{K \rightarrow \infty} \mathcal{L}_K(q, \boldsymbol{\theta})$  if  $\frac{p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z}|\mathbf{x})}$  is bounded.

## Proof of 1.

$$\begin{aligned}
 \mathcal{L}_K(q, \boldsymbol{\theta}) &= \mathbb{E}_{\mathbf{z}_1, \dots, \mathbf{z}_K} \log \left( \frac{1}{K} \sum_{k=1}^K \frac{p(\mathbf{x}, \mathbf{z}_k|\boldsymbol{\theta})}{q(\mathbf{z}_k|\mathbf{x})} \right) = \\
 &= \mathbb{E}_{\mathbf{z}_1, \dots, \mathbf{z}_K} \log \mathbb{E}_{k_1, \dots, k_M} \left( \frac{1}{M} \sum_{m=1}^M \frac{p(\mathbf{x}, \mathbf{z}_{k_m}|\boldsymbol{\theta})}{q(\mathbf{z}_{k_m}|\mathbf{x})} \right) \geq \\
 &\geq \mathbb{E}_{\mathbf{z}_1, \dots, \mathbf{z}_K} \mathbb{E}_{k_1, \dots, k_M} \log \left( \frac{1}{M} \sum_{m=1}^M \frac{p(\mathbf{x}, \mathbf{z}_{k_m}|\boldsymbol{\theta})}{q(\mathbf{z}_{k_m}|\mathbf{x})} \right) = \\
 &= \mathbb{E}_{\mathbf{z}_1, \dots, \mathbf{z}_M} \log \left( \frac{1}{M} \sum_{m=1}^M \frac{p(\mathbf{x}, \mathbf{z}_m|\boldsymbol{\theta})}{q(\mathbf{z}_m|\mathbf{x})} \right) = \mathcal{L}_M(q, \boldsymbol{\theta})
 \end{aligned}$$

## Theorem

1.  $\log p(\mathbf{x}) \geq \mathcal{L}_K(q, \boldsymbol{\theta}) \geq \mathcal{L}_M(q, \boldsymbol{\theta})$ , for  $K \geq M$ ;
2.  $\log p(\mathbf{x}) = \lim_{K \rightarrow \infty} \mathcal{L}_K(q, \boldsymbol{\theta})$  if  $\frac{p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z}|\mathbf{x})}$  is bounded.

## Proof of 2.

Consider r.v.  $\xi_K = \frac{1}{K} \sum_{k=1}^K \frac{p(\mathbf{x}, \mathbf{z}_k|\boldsymbol{\theta})}{q(\mathbf{z}_k|\mathbf{x})}$ .

If summands are bounded, then (from the strong law of large numbers)

$$\xi_K \xrightarrow[K \rightarrow \infty]{a.s.} \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \frac{p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z}|\mathbf{x})} = p(\mathbf{x}|\boldsymbol{\theta}).$$

Hence  $\mathcal{L}_K(q, \boldsymbol{\theta}) = \mathbb{E} \log \xi_K$  converges to  $\log p(\mathbf{x}|\boldsymbol{\theta})$  as  $K \rightarrow \infty$ .

$$\log p(\mathbf{x}|\boldsymbol{\theta}) \geq \mathcal{L}_K(q, \boldsymbol{\theta}) \geq \mathcal{L}(q, \boldsymbol{\theta})$$

If  $K > 1$  the bound could be tighter.

$$\mathcal{L}(q, \boldsymbol{\theta}) = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \log \frac{p(\mathbf{x}, \mathbf{z}|\boldsymbol{\theta})}{q(\mathbf{z}|\mathbf{x})};$$

$$\mathcal{L}_K(q, \boldsymbol{\theta}) = \mathbb{E}_{\mathbf{z}_1, \dots, \mathbf{z}_K \sim q(\mathbf{z}|\mathbf{x})} \log \left( \frac{1}{K} \sum_{k=1}^K \frac{p(\mathbf{x}, \mathbf{z}_k|\boldsymbol{\theta})}{q(\mathbf{z}_k|\mathbf{x})} \right).$$

- ▶  $\mathcal{L}_1(q, \boldsymbol{\theta}) = \mathcal{L}(q, \boldsymbol{\theta})$ ;
- ▶  $\mathcal{L}_\infty(q, \boldsymbol{\theta}) = \log p(\mathbf{x}|\boldsymbol{\theta})$ .

Which  $q(\mathbf{z}|\mathbf{x})$  gives  $\mathcal{L}(q, \boldsymbol{\theta}) = \log p(\mathbf{x}|\boldsymbol{\theta})$ ?

Which  $q(\mathbf{z}|\mathbf{x})$  gives  $\mathcal{L}(q, \boldsymbol{\theta}) = \mathcal{L}_K(q, \boldsymbol{\theta})$ ?



## Theorem

The VAE objective is equal to IWAE objective

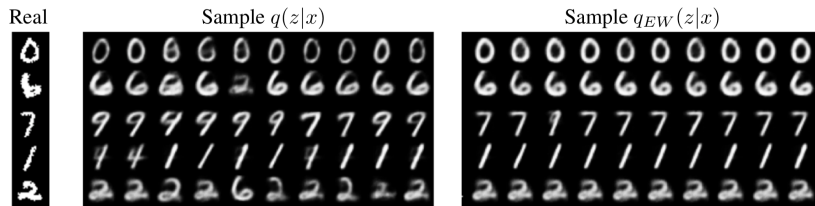
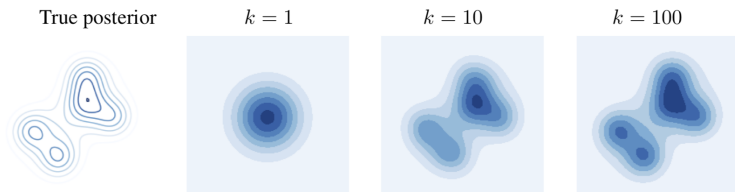
$$\mathcal{L}(q_{EW}, \theta) = \mathcal{L}_K(q, \theta)$$

for the following variational distribution

$$q_{EW}(\mathbf{z}|\mathbf{x}) = \mathbb{E}_{\mathbf{z}_2, \dots, \mathbf{z}_K \sim q(\mathbf{z}|\mathbf{x})} q_{IW}(\mathbf{z}|\mathbf{x}, \mathbf{z}_{2:K}),$$

where

$$q_{IW}(\mathbf{z}|\mathbf{x}, \mathbf{z}_{2:K}) = \frac{\frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z}|\mathbf{x})}}{\frac{1}{K} \sum_{k=1}^K \frac{p(\mathbf{x}, \mathbf{z}_k)}{q(\mathbf{z}_k|\mathbf{x})}} q(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{x}, \mathbf{z})}{\frac{1}{K} \left( \frac{p(\mathbf{x}, \mathbf{z})}{q(\mathbf{z}|\mathbf{x})} + \sum_{k=2}^K \frac{p(\mathbf{x}, \mathbf{z}_k)}{q(\mathbf{z}_k|\mathbf{x})} \right)}.$$



How to determine whether all VAE latent variables are informative?

$$A_i = \text{cov}_{\mathbf{x}} \left( \mathbb{E}_{q(z_i|\mathbf{x})}[z_i] \right) > 0.01 \quad \Leftrightarrow \quad z_i \text{ is active}$$

# stoch. layers	$k$	MNIST				OMNIGLOT			
		VAE		IWAE		VAE		IWAE	
		NLL	active units	NLL	active units	NLL	active units	NLL	active units
1	1	86.76	19	86.76	19	108.11	28	108.11	28
	5	86.47	20	85.54	22	107.62	28	106.12	34
	50	86.35	20	84.78	25	107.80	28	104.67	41
2	1	85.33	16+5	85.33	16+5	107.58	28+4	107.56	30+5
	5	85.01	17+5	83.89	21+5	106.31	30+5	104.79	38+6
	50	84.78	17+5	82.90	26+7	106.30	30+5	103.38	44+7

<https://arxiv.org/pdf/1509.00519.pdf>

## ELBO interpretations

- ▶ Evidence minus posterior KL

$$\mathcal{L}(q, \theta) = \log p(\mathbf{X}|\theta) + KL(q(\mathbf{Z}|\mathbf{X})||p(\mathbf{Z}|\mathbf{X}, \theta)).$$

- ▶ Average negative energy plus entropy

$$\mathcal{L}(q, \theta) = \mathbb{E}_{q(\mathbf{Z}|\mathbf{X})} p(\mathbf{X}, \mathbf{Z}|\theta) + \mathbb{H}[q(\mathbf{Z}|\mathbf{X})].$$

- ▶ Average term-by-term reconstruction minus KL to prior

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

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

## Theorem

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

where  $i$  is treated as random variable:

$$q(i, \mathbf{z}) = q(i)q(\mathbf{z}|i); \quad p(i, \mathbf{z}) = p(i)p(\mathbf{z}); \quad q(i) = p(i) = \frac{1}{n}; \quad q(\mathbf{z}|i) = q(\mathbf{z}|\mathbf{x}_i).$$

$$q(\mathbf{z}) = \sum_{i=1}^n q(i, \mathbf{z}) = \frac{1}{n} \sum_{i=1}^n q(\mathbf{z}|\mathbf{x}_i); \quad \mathbb{I}_{q(i, \mathbf{z})}[i, \mathbf{z}] = \mathbb{E}_{q(i, \mathbf{z})} \log \frac{q(i, \mathbf{z})}{q(i)q(\mathbf{z})}.$$

## Theorem

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

## Proof

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n KL(q(\mathbf{z}_i | \mathbf{x}_i) || p(\mathbf{z}_i)) &= \sum_{i=1}^n \int q(i) q(\mathbf{z} | i) \log \frac{q(\mathbf{z} | i)}{p(\mathbf{z})} d\mathbf{z} = \\ &= \sum_{i=1}^n \int q(i, \mathbf{z}) \log \frac{q(i, \mathbf{z})}{p(\mathbf{z}) p(i)} d\mathbf{z} = \int \sum_{i=1}^n q(i, \mathbf{z}) \log \frac{q(\mathbf{z}) q(i | \mathbf{z})}{p(\mathbf{z}) p(i)} d\mathbf{z} = \\ &= \int q(\mathbf{z}) \log \frac{q(\mathbf{z})}{p(\mathbf{z})} d\mathbf{z} + \int \sum_{i=1}^n q(i | \mathbf{z}) q(\mathbf{z}) \log \frac{q(i | \mathbf{z})}{p(i)} d\mathbf{z} = \\ &= KL(q(\mathbf{z}) || p(\mathbf{z})) - \mathbb{E}_{q(\mathbf{z})} \mathbb{H} [q(i | \mathbf{z})] + \log N. \end{aligned}$$

## Theorem

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

## Proof (continued)

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

$$\begin{aligned} \mathbb{I}_{q(i, \mathbf{z})}[i, \mathbf{z}] &= \mathbb{E}_{q(i, \mathbf{z})} \log \frac{q(i, \mathbf{z})}{q(i)q(\mathbf{z})} = \mathbb{E}_{q(\mathbf{z})} \mathbb{E}_{q(i | \mathbf{z})} \log \frac{q(i | \mathbf{z})q(\mathbf{z})}{q(i)q(\mathbf{z})} = \\ &= \mathbb{E}_{q(\mathbf{z})} \mathbb{E}_{q(i | \mathbf{z})} \log \frac{q(i | \mathbf{z})}{q(i)} = -\mathbb{E}_{q(\mathbf{z})} \mathbb{H}[q(i | \mathbf{z})] + \log N. \end{aligned}$$

# ELBO surgery, 2016

## Theorem

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

## ELBO revisiting

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



# ELBO surgery, 2016

## ELBO revisiting

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

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

	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

$$\log N \approx 11.0021$$

## ELBO revisiting

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

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

- ▶ SG:  $p(\mathbf{z}) = \mathcal{N}(0, I) \Rightarrow$  over-regularization;
- ▶ MoG:  $p(\mathbf{z} | \lambda) = \frac{1}{K} \sum_{k=1}^K \mathcal{N}(\mu_k, \sigma_k^2) \Rightarrow (*), (**);$
- ▶  $p(\mathbf{z}) = q(\mathbf{z}) = \frac{1}{n} \sum_{i=1}^n q(\mathbf{z} | \mathbf{x}_i) \Rightarrow$  overfitting and highly expensive.

---

(\*) <https://arxiv.org/abs/1611.02648>

(\*\*) <https://pdfs.semanticscholar.org/f6fe/5e8e25994c188ba6a124462e2cc55f2c5a67.pdf>

## Variational Mixture of posteriors

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

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

---

<https://arxiv.org/pdf/1705.07120.pdf>

# VampPrior, 2017

Do we equally need the multimodal prior?

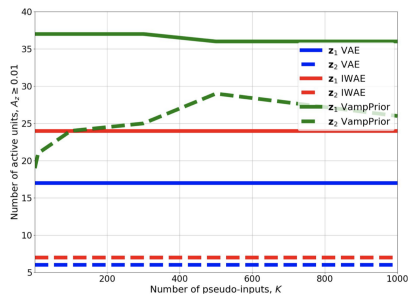
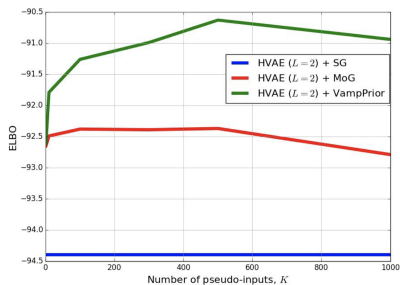
Is it beneficial to couple the prior with the variational posterior or MoG is enough?

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 <i>data</i>	-85.71
HVAE ( $L = 2$ ) + VAMPprior	<b>-83.19</b>
AVB + AC ( $L = 1$ ) [28]	-80.20
VLAE [7]	<b>-79.03</b>
VAE + IAF [18]	-79.88
CONVHVAE ( $L = 2$ ) + VAMPprior	-81.09
PIXELHVAE ( $L = 2$ ) + VAMPprior	-79.78

---

<https://arxiv.org/pdf/1705.07120.pdf>

# VampPrior, 2017

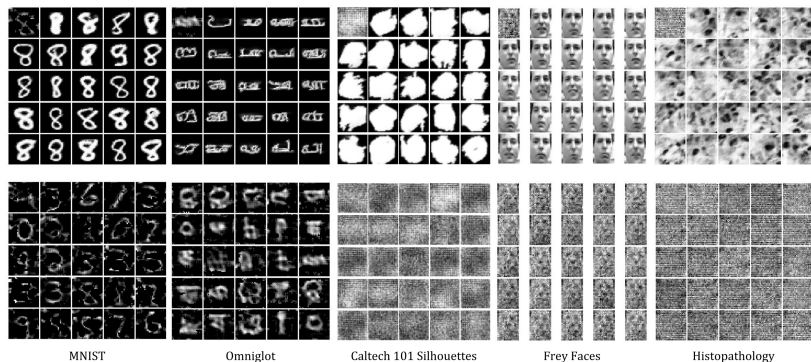


<https://arxiv.org/pdf/1705.07120.pdf>

# VampPrior, 2017

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

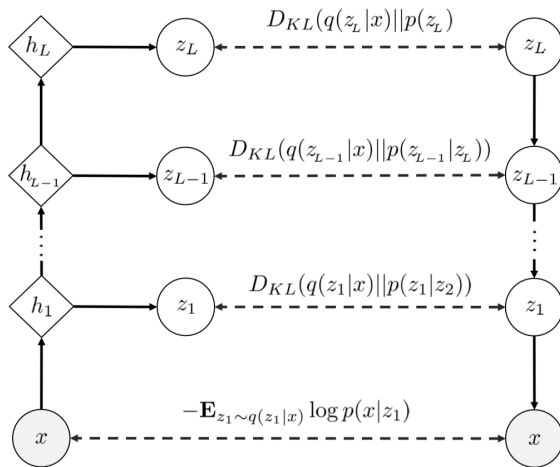
**Bottom row:** Images represent a subset of trained pseudo-inputs for different datasets.



# References

- ▶ **IWAE: Importance Weighted Autoencoders**  
<https://arxiv.org/pdf/1509.00519.pdf>  
**Summary:** Propose the version of ELBO which is tighter to the log-density. Sampling of  $k$  objects from latent space instead of one. The more samples we samples the more tight the gap is. In the limit we will get the true likelihood. The gradient is given by importance sampling reweighting. Analyze the number of active latent units.
- ▶ **Reinterpreting Importance-Weighted Autoencoders**  
<https://arxiv.org/abs/1704.02916>  
**Summary:** IWAE maximizes a tighter ELBO than VAE ELBO. This ELBO is also related to some implicit variational distribution. The analytical form of this distribution is given and visualized.
- ▶ **ELBO surgery: yet another way to carve up the variational evidence lower bound**  
<http://approximateinference.org/accepted/HoffmanJohnson2016.pdf>  
**Summary:** Propose the decomposition of standard ELBO into 3 terms. The prior distribution should be close to average posterior. Show empirically that weak prior has a significant impact on ELBO value.
- ▶ **VAE with a VampPrior**  
<https://arxiv.org/pdf/1705.07120.pdf>  
**Summary:** Variational Mixture of Posteriors prior is introduced. The VampPrior components are given by variational posteriors conditioned on learnable pseudo-inputs. Prior is extended to a two layer hierarchical model with a coupled prior and posterior, it learns significantly better models. The model avoids the local optima issues related to useless latent dimensions that plague VAEs. The prior is compared with standard gaussian and mixture of gaussians.

## Hierarchical VAE





## 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

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

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

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

Autoregressive decoder

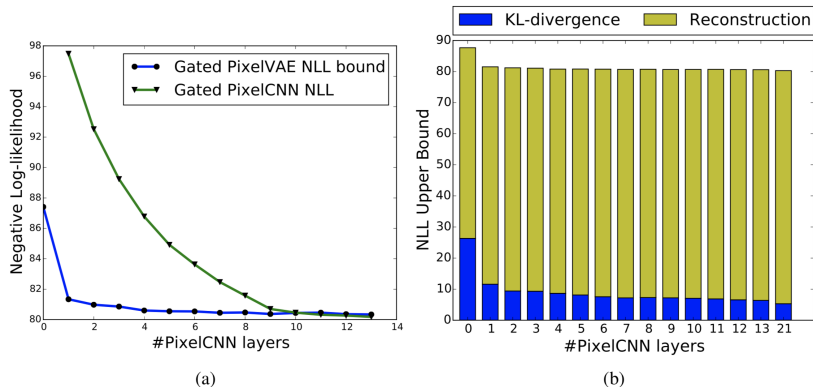
$$p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) = \prod_{i=1}^n p(x_i|\mathbf{x}_{1:i-1}, \mathbf{z}, \boldsymbol{\theta})$$

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

---

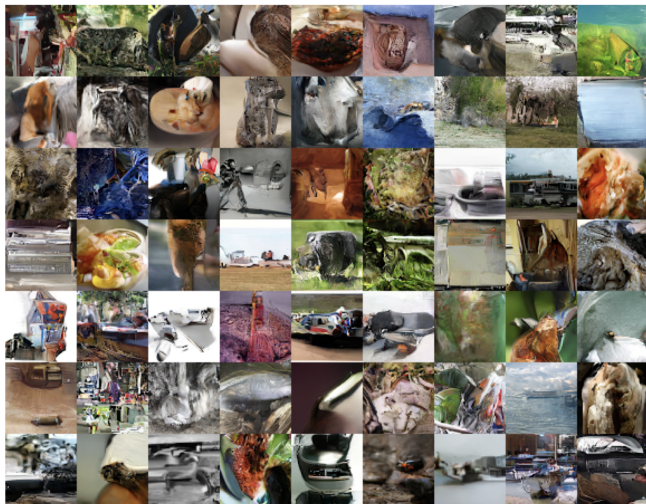
<https://arxiv.org/pdf/1611.05013.pdf>

# PixelVAE, 2016



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

# PixelVAE, 2016



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

# PixelVAE, 2016

<b>Model</b>	<b>NLL Validation (Train)</b>
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)	= <b>3.57</b> (3.48)
Hierarchical PixelVAE	$\leq$ 3.66 (3.59)

# 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  $p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) = \pi(\mathbf{x})$ .

# 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  $\lambda$  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

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

- ▶  $\text{WindowAround}(i)$  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

## Theorem

AF prior is equivalent to IAF posterior.

## Proof

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

- ▶ 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.

# 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	Method	bits/dim $\leq$
<i>Results with tractable likelihood models:</i>			
Normalizing flows (Rezende & Mohamed, 2015)	85.10	Uniform distribution [1]	8.00
DRAW (Gregor et al., 2015)	< 80.97	Multivariate Gaussian [1]	4.70
Discrete VAE (Rolfe, 2016)	81.01	NICE [2]	4.48
PixelRNN (van den Oord et al., 2016a)	79.20	Deep GMMs [3]	4.00
IAF VAE (Kingma et al., 2016)	79.88	Real NVP [4]	3.49
AF VAE	79.30	PixelCNN [1]	3.14
VLAE	<b>79.03</b>	Gated PixelCNN [5]	3.03
		PixelRNN [1]	3.00
		PixelCNN++ [6]	<b>2.92</b>
<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	<b>2.95</b>

<https://arxiv.org/abs/1606.04934>

# 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>

## 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  $\mathbf{v}$ .

## Constrained optimization

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

## Objective

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

What 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.

## 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.

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

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

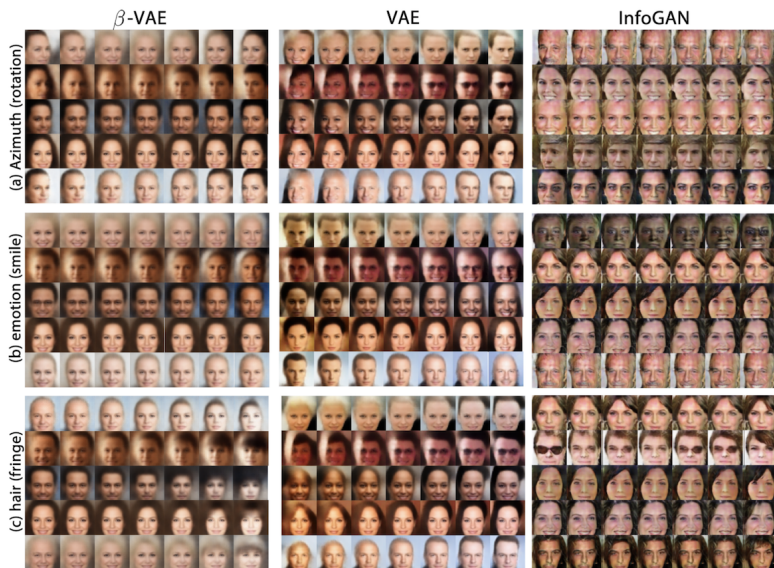
$$q(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mu(\mathbf{x})|\sigma^2(\mathbf{x})); \quad \mathbf{z}_{li} = \mu(\mathbf{x}_{li}); \quad \mathbf{z}_{lj} = \mu(\mathbf{x}_{lj}).$$

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

---

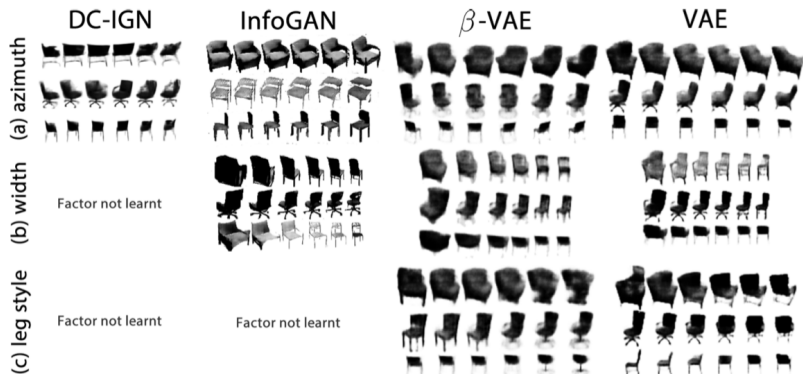
<https://openreview.net/references/pdf?id=Sy2fzU9gl>

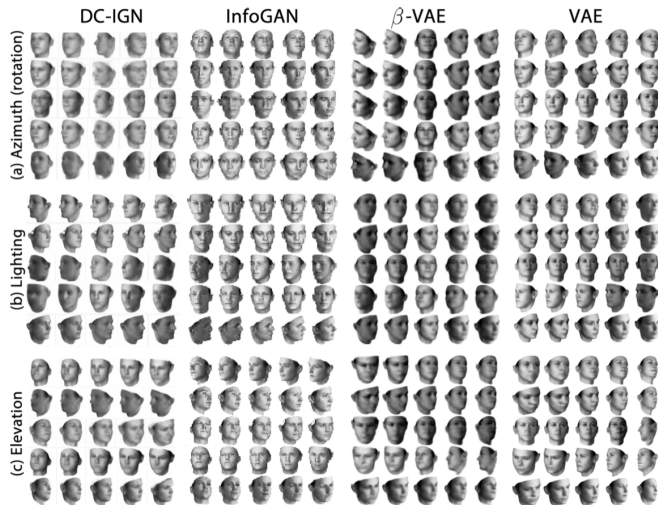
# $\beta$ -VAE, 2017





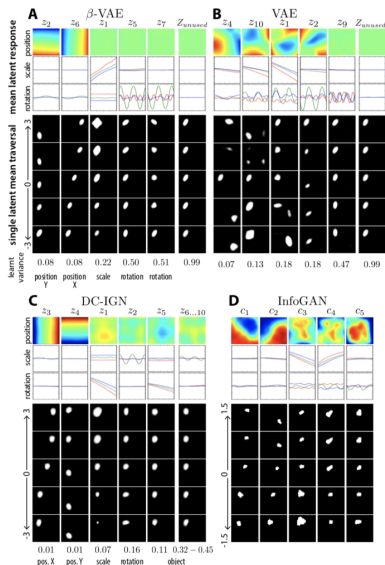
# $\beta$ -VAE, 2017



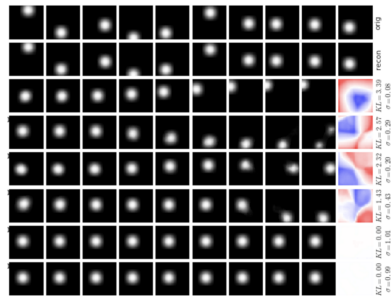


# $\beta$ -VAE, 2017

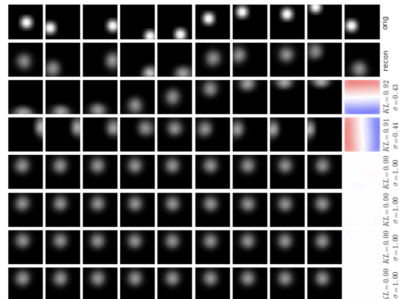
Model	Disentanglement metric score
Ground truth	100%
Raw pixels	$45.75 \pm 0.8\%$
PCA	$84.9 \pm 0.4\%$
ICA	$42.03 \pm 10.6\%$
DC-IGN	<b><math>99.3 \pm 0.1\%</math></b>
InfoGAN	$73.5 \pm 0.9\%$
VAE untrained	$44.14 \pm 2.5\%$
VAE	$61.58 \pm 0.5\%$
$\beta$ -VAE	<b><math>99.23 \pm 0.1\%</math></b>



$\beta = 1$



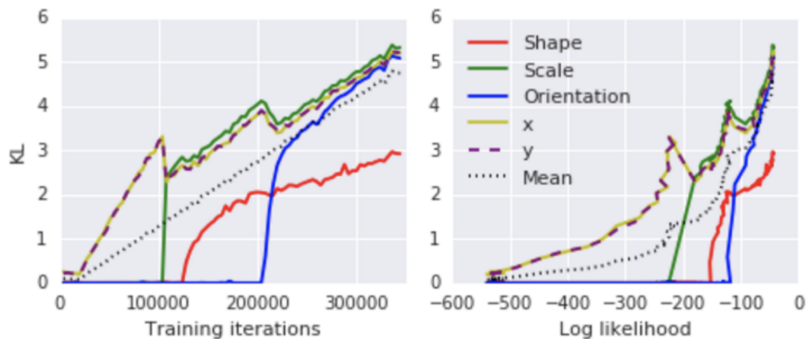
$\beta = 150$



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

## Controlled encoding capacity

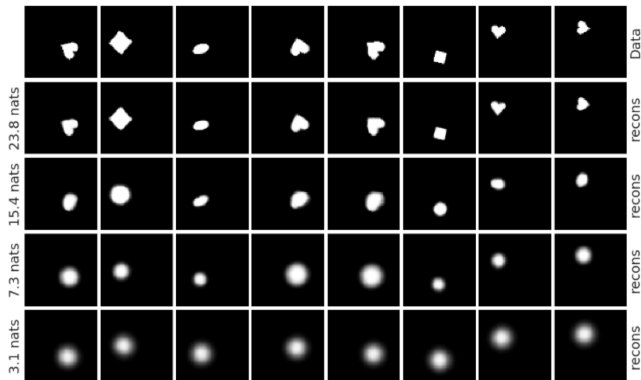
$$\mathcal{L}(q, \theta, \beta) = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \log p(\mathbf{x}|\mathbf{z}, \theta) - |KL(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) - C|.$$



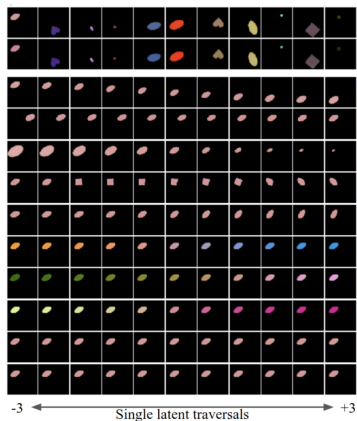
# $\beta$ -VAE, 2018

## Controlled encoding capacity

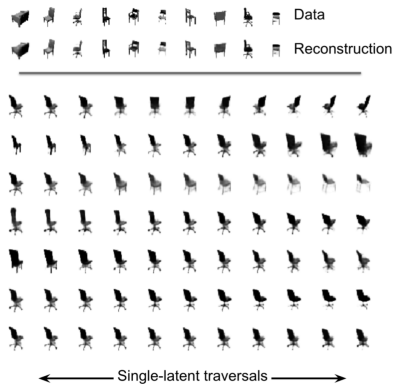
$$\mathcal{L}(q, \theta, \beta) = \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \log p(\mathbf{x}|\mathbf{z}, \theta) - |KL(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) - C|.$$



(a) Coloured dSprites



(b) 3D Chairs



# 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  $\beta$ -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.