

Deep Generative Models

Lecture 8

Roman Isachenko



Ozon Masters

Spring, 2021

Recap of previous lecture

Images are discrete data, flow is a continuous model. We need to convert a discrete data distribution to a continuous one.

Uniform dequantization bound

$\mathbf{x} \sim \text{Categorical}(\boldsymbol{\pi}), \quad \mathbf{u} \sim U[0, 1], \quad \mathbf{y} = \mathbf{x} + \mathbf{u} \sim \text{Continuous}$

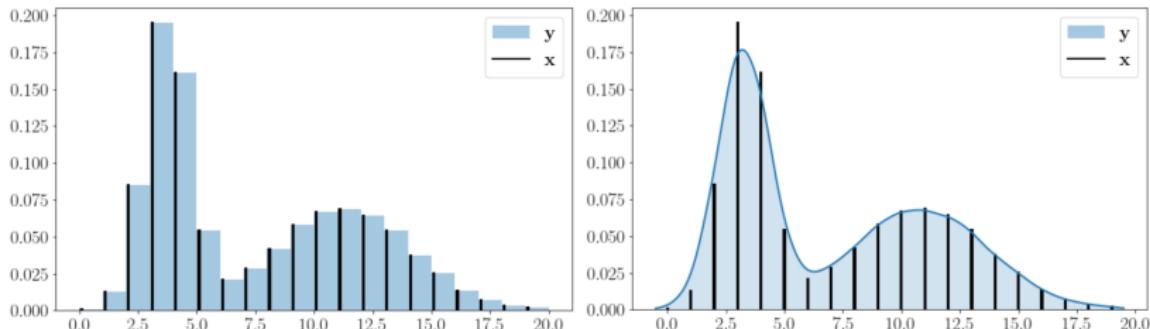
$$\log P(\mathbf{x}|\theta) \geq \int_{U[0,1]} \log p(\mathbf{x} + \mathbf{u}|\theta) d\mathbf{u}.$$

Variational dequantization bound

Introduce variational dequantization noise distribution $q(\mathbf{u}|\mathbf{x})$ and treat it as an approximate posterior.

$$\log P(\mathbf{x}|\theta) \geq \int q(\mathbf{u}|\mathbf{x}) \log \frac{p(\mathbf{x} + \mathbf{u}|\theta)}{q(\mathbf{u}|\mathbf{x})} d\mathbf{u} = \mathcal{L}(q, \theta).$$

Recap of previous lecture



Flow model for dequantization

$$q(\mathbf{u}|\mathbf{x}) = p(h^{-1}(\mathbf{u}, \phi)) \cdot \left| \det \frac{\partial h^{-1}(\mathbf{u}, \phi)}{\partial \mathbf{u}} \right|.$$

Variational dequantization bound

$$\mathcal{L}(q, \theta) = \int q(\mathbf{u}|\mathbf{x}) \log \frac{p(\mathbf{x} + \mathbf{u}|\theta)}{q(\mathbf{u}|\mathbf{x})} d\mathbf{u}.$$

Recap of previous lecture

VAE objective

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

IWAE objective

$$\mathcal{L}_K(q, \theta) = \mathbb{E}_{z_1, \dots, z_K \sim q(z|x)} \log \left(\frac{1}{K} \sum_{k=1}^K \frac{p(\mathbf{x}, z_k|\theta)}{q(z_k|x)} \right) \rightarrow \max_{q,\theta}.$$

Theorem

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

Recap of previous lecture

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

Theorem

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

- ▶ $q(\mathbf{z}) = \frac{1}{n} \sum_{i=1}^n q(\mathbf{z}|\mathbf{x}_i)$ – **aggregated** posterior distribution.
- ▶ $\mathbb{I}_q[\mathbf{x}, \mathbf{z}]$ – mutual information between \mathbf{x} and \mathbf{z} under empirical data distribution and distribution $q(\mathbf{z}|\mathbf{x})$.
- ▶ First term pushes $q(\mathbf{z})$ towards the prior $p(\mathbf{z})$.
- ▶ Second term reduces the amount of information about \mathbf{x} stored in \mathbf{z} .

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

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

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

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

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

The optimal prior distribution $p(\mathbf{z})$ is aggregated posterior $q(\mathbf{z})$.
How to choose the optimal $p(\mathbf{z})$?

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

VampPrior

Optimal prior

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

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

Variational Mixture of posteriors

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

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

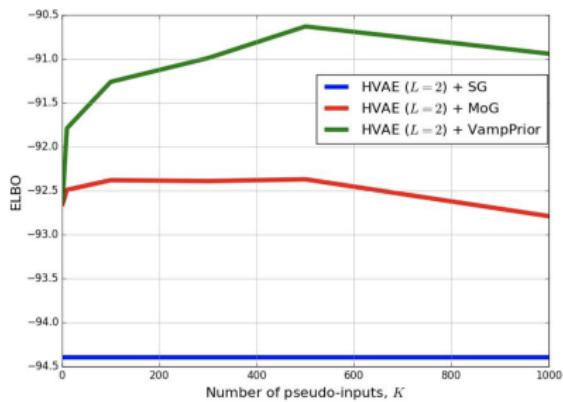
- ▶ Multimodal \Rightarrow prevents over-regularization;.
- ▶ $K \ll n \Rightarrow$ prevents from potential overfitting + less expensive to train.
- ▶ Pseudo-inputs are prior hyperparameters \Rightarrow connection to the Empirical Bayes.

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

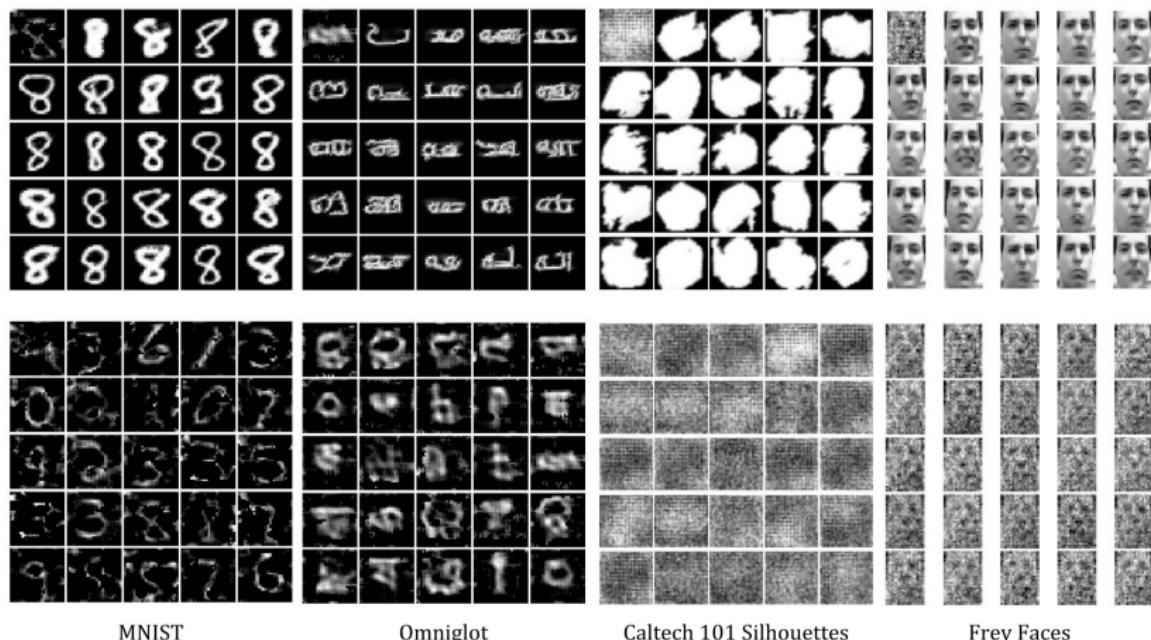
MODEL	LL
VAE ($L = 1$) + NF [32]	-85.10
VAE ($L = 2$) [6]	-87.86
IWAE ($L = 2$) [6]	-85.32
HVAE ($L = 2$) + SG	-85.89
HVAE ($L = 2$) + MoG	-85.07
HVAE ($L = 2$) + VAMPRIOR data	-85.71
HVAE ($L = 2$) + VAMPRIOR	-83.19
AVB + AC ($L = 1$) [28]	-80.20
VLAЕ [7]	-79.03
VAE + IAF [18]	-79.88
convHVAE ($L = 2$) + VAMPRIOR	-81.09
PIXELHVAE ($L = 2$) + VAMPRIOR	-79.78



VampPrior

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

Bottom row: pseudo-inputs for different datasets.



Flow prior in VAE

ELBO revisiting

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

VampPrior

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

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

Autoregressive flow prior

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

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

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 ϵ .

Proof

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

Flows in VAE

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

Flow prior in VAE

Autoregressive flow prior

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

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

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.

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

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} becomes irrelevant.

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

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

LVM

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

- ▶ 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

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

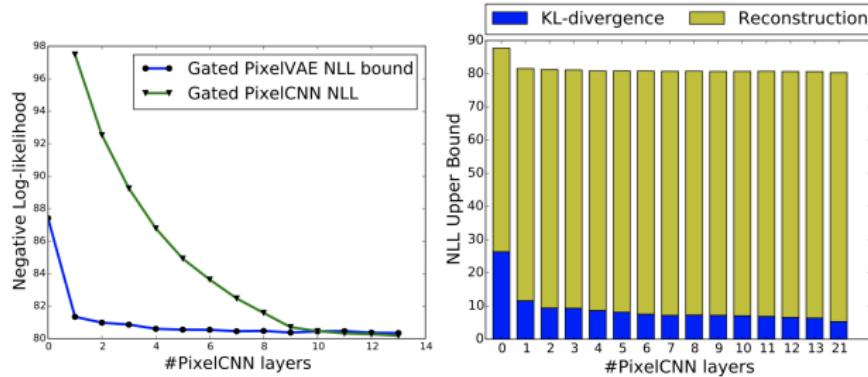
PixelVAE

Autoregressive decoder

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

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

MNIST results



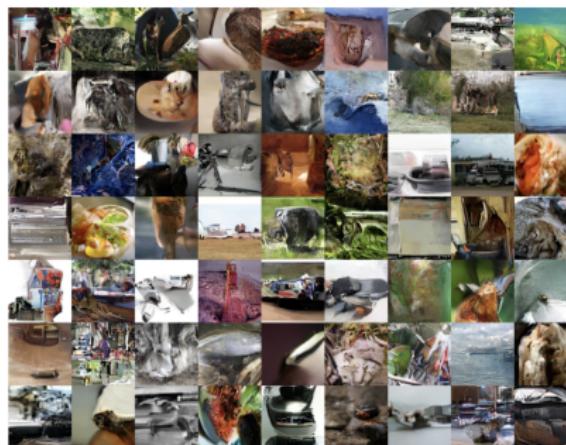
PixelVAE

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

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

- ▶ Powerful decoder $p(\mathbf{x}|\mathbf{z}, \theta)$ 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} ?

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

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

KL annealing

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

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.

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.

Disentanglement learning

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

ELBO 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 $\beta = 1$?

Constrained optimization

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

Hypothesis

We are able to learn disentangled representations of the independent factors v by setting a stronger constraint with $\beta > 1$.

Note: It leads to poorer reconstructions and a loss of high frequency details.

β -VAE

β -VAE



(a) Azimuth (rotation)

VAE



(a) Azimuth (rotation)

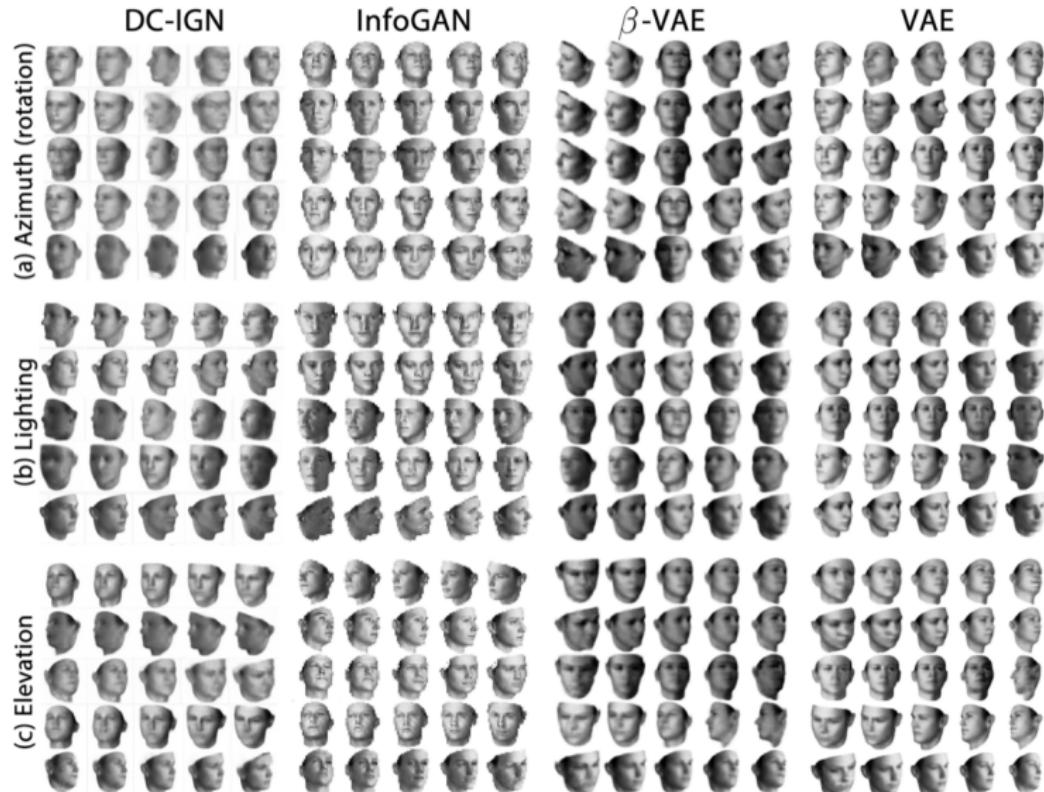


(b) emotion (smile)



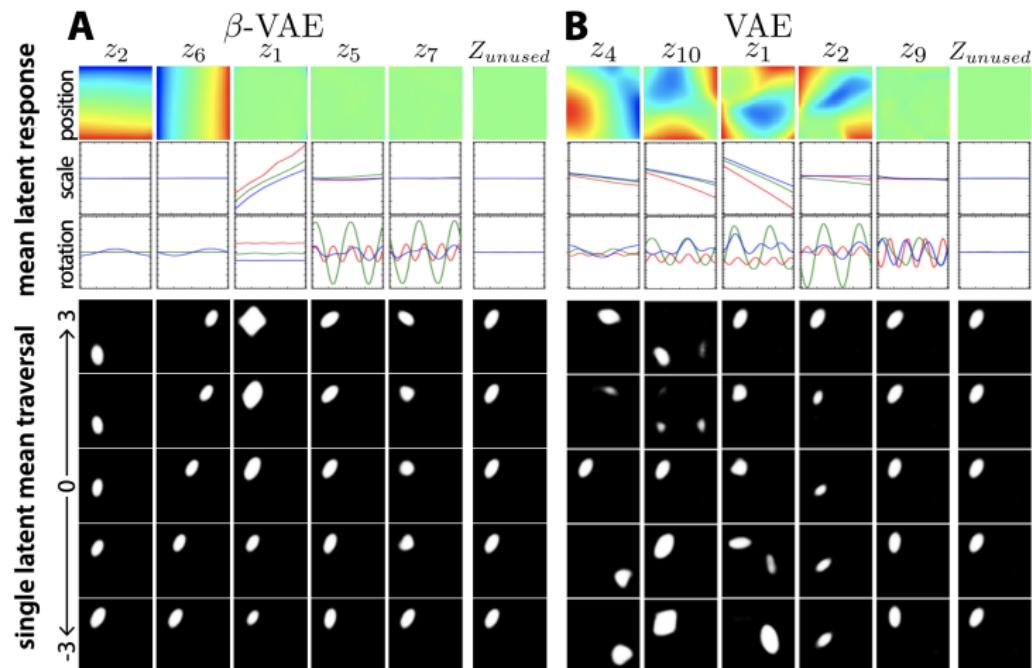
Higgins I. et al. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework, 2017

β -VAE



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

β -VAE



β -VAE

ELBO

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

ELBO surgery

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

Minimization of MI

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

DIP-VAE

Disentangled aggregated variational posterior

$$q(\mathbf{z}) = \frac{1}{n} \sum_{i=1}^n q(\mathbf{z}|\mathbf{x}) = \prod_{j=1}^d q(z_j)$$

DIP-VAE Objective

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

DIP-VAE

$$\mathcal{L}_{\text{DIP}}(q, \theta) = \frac{1}{n} \sum_{i=1}^n \mathcal{L}_i(q, \theta) - \lambda \cdot \underbrace{KL(q(\mathbf{z}) || p(\mathbf{z}))}_{\text{intractable}}$$

Let match the moments of $q(\mathbf{z})$ and $p(\mathbf{z})$:

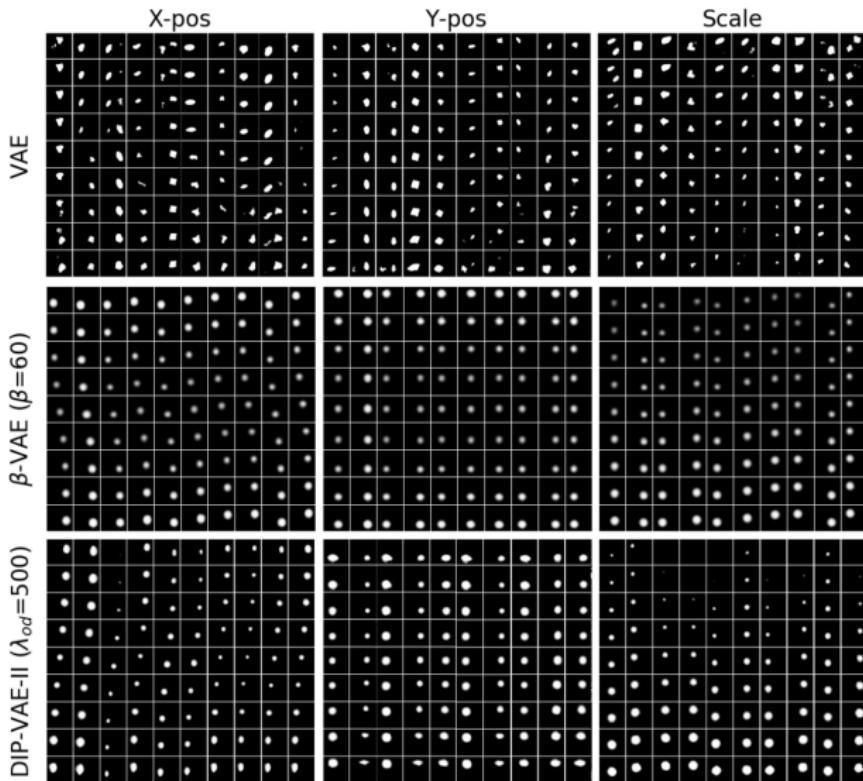
$$\text{cov}_{q(\mathbf{z})}(\mathbf{z}) = \mathbb{E}_{q(\mathbf{z})} \left[(\mathbf{z} - \mathbb{E}_{q(\mathbf{z})}(\mathbf{z})) (\mathbf{z} - \mathbb{E}_{q(\mathbf{z})}(\mathbf{z}))^T \right]$$

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

Objective

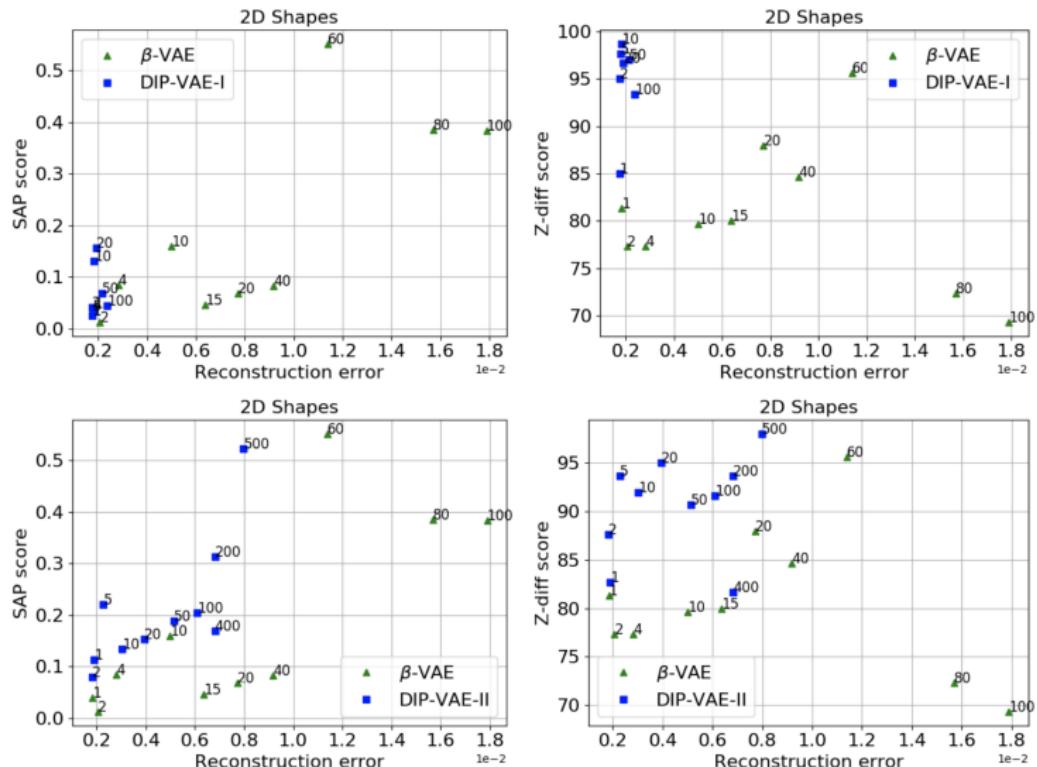
$$\begin{aligned} \max_{q, \theta} & \left[\frac{1}{n} \sum_{i=1}^n \mathcal{L}_i(q, \theta) - \right. \\ & \left. - \lambda_1 \sum_{i \neq j} [\text{cov}_{q(\mathbf{z})}(\mathbf{z})]_{ij}^2 - \lambda_2 \sum_i ([\text{cov}_{q(\mathbf{z})}(\mathbf{z})]_{ii} - 1)^2 \right] \end{aligned}$$

DIP-VAE



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

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

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

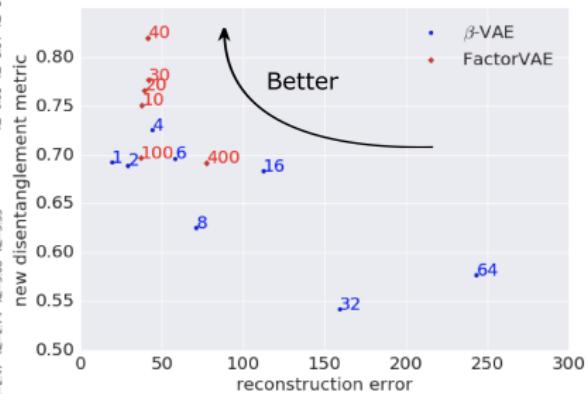
VAE and GAN are trained simultaneously.

FactorVAE

β -VAE ($\beta = 8$)



FactorVAE ($\gamma = 10$)



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.