

Deep Generative Models

Lecture 8

Roman Isachenko

Moscow Institute of Physics and Technology

Autumn, 2021

Flows in VAE posterior

Apply a sequence of transformations to the random variable

$$\mathbf{z}_0 \sim q(\mathbf{z}|\mathbf{x}, \phi) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_\phi(\mathbf{x}), \boldsymbol{\sigma}_\phi^2(\mathbf{x})).$$

Let $q(\mathbf{z}|\mathbf{x}, \phi)$ (VAE encoder) be a base distribution for a flow model.

Flow model in latent space

$$\log q(\mathbf{z}^*|\mathbf{x}, \phi, \boldsymbol{\lambda}) = \log q(\mathbf{z}|\mathbf{x}, \phi) + \log \det \left| \frac{\partial f(\mathbf{z}, \boldsymbol{\lambda})}{\partial \mathbf{z}} \right|$$
$$\mathbf{z}^* = f(\mathbf{z}, \boldsymbol{\lambda}) = g^{-1}(\mathbf{z}, \boldsymbol{\lambda})$$

Here $f(\mathbf{z}, \boldsymbol{\lambda})$ is a flow model (e.g. stack of planar/coupling layers) parameterized by $\boldsymbol{\lambda}$.

Let use $q(\mathbf{z}^*|\mathbf{x}, \phi, \boldsymbol{\lambda})$ as a variational distribution. Here ϕ – encoder parameters, $\boldsymbol{\lambda}$ – flow parameters.

Flows-based VAE posterior

- ▶ Encoder outputs base distribution $q(\mathbf{z}|\mathbf{x}, \phi)$.
- ▶ Flow model $\mathbf{z}^* = f(\mathbf{z}, \lambda)$ transforms the base distribution $q(\mathbf{z}|\mathbf{x}, \phi)$ to the distribution $q(\mathbf{z}^*|\mathbf{x}, \phi, \lambda)$.
- ▶ Distribution $q(\mathbf{z}^*|\mathbf{x}, \phi, \lambda)$ is used as a variational distribution for ELBO maximization.

Flow model in latent space

$$\log q(\mathbf{z}^*|\mathbf{x}, \phi, \lambda) = \log q(\mathbf{z}|\mathbf{x}, \phi) + \log \det \left| \frac{\partial f(\mathbf{z}, \lambda)}{\partial \mathbf{z}} \right|$$

ELBO with flow-based VAE posterior

$$\begin{aligned}\mathcal{L}(\phi, \theta, \lambda) &= \mathbb{E}_{q(\mathbf{z}^*|\mathbf{x}, \phi, \lambda)} [\log p(\mathbf{x}, \mathbf{z}^*|\theta) - \log q(\mathbf{z}^*|\mathbf{x}, \phi, \lambda)] \\ &= \mathbb{E}_{q(\mathbf{z}^*|\mathbf{x}, \phi, \lambda)} \log p(\mathbf{x}|\mathbf{z}^*, \theta) - KL(q(\mathbf{z}^*|\mathbf{x}, \phi, \lambda) || p(\mathbf{z}^*)).\end{aligned}$$

The second term in ELBO is reverse KL divergence. Planar flows was originally proposed for variational inference in VAE.

Flows-based VAE posterior

Flow model in latent space

$$\log q(\mathbf{z}^* | \mathbf{x}, \phi, \lambda) = \log q(\mathbf{z} | \mathbf{x}, \phi) + \log \det \left| \frac{\partial f(\mathbf{z}, \lambda)}{\partial \mathbf{z}} \right|$$

ELBO objective

$$\begin{aligned}\mathcal{L}(\phi, \theta, \lambda) &= \mathbb{E}_{q(\mathbf{z}^* | \mathbf{x}, \phi, \lambda)} [\log p(\mathbf{x}, \mathbf{z}^* | \theta) - \log q(\mathbf{z}^* | \mathbf{x}, \phi, \lambda)] = \\ &= \mathbb{E}_{q(\mathbf{z} | \mathbf{x}, \phi)} [\log p(\mathbf{x}, \mathbf{z}^* | \theta) - \log q(\mathbf{z}^* | \mathbf{x}, \phi, \lambda)] \Big|_{\mathbf{z}^* = f(\mathbf{z}, \lambda)} = \\ &= \mathbb{E}_{q(\mathbf{z} | \mathbf{x}, \phi)} \left[\log p(\mathbf{x}, \mathbf{z}^* | \theta) - \log q(\mathbf{z} | \mathbf{x}, \phi) + \log \left| \det \left(\frac{\partial f(\mathbf{z}, \lambda)}{\partial \mathbf{z}} \right) \right| \right].\end{aligned}$$

- ▶ Obtain samples \mathbf{z} from the encoder $q(\mathbf{z} | \mathbf{x}, \phi)$.
- ▶ Apply flow model $\mathbf{z}^* = f(\mathbf{z}, \lambda)$.
- ▶ Compute likelihood for \mathbf{z}^* using the decoder, base distribution for \mathbf{z}^* and the Jacobian.

Inverse autoregressive flow (IAF)

$$\mathbf{x} = g(\mathbf{z}, \boldsymbol{\theta}) \quad \Rightarrow \quad x_i = \tilde{\sigma}_i(\mathbf{z}_{1:i-1}) \cdot z_i + \tilde{\mu}_i(\mathbf{z}_{1:i-1}).$$

$$\mathbf{z} = f(\mathbf{x}, \boldsymbol{\theta}) \quad \Rightarrow \quad z_i = (x_i - \tilde{\mu}_i(\mathbf{z}_{1:i-1})) \cdot \frac{1}{\tilde{\sigma}_i(\mathbf{z}_{1:i-1})}.$$

Reverse KL for flow model

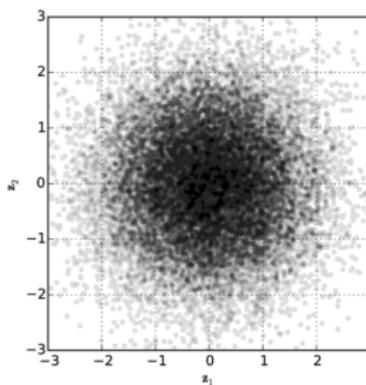
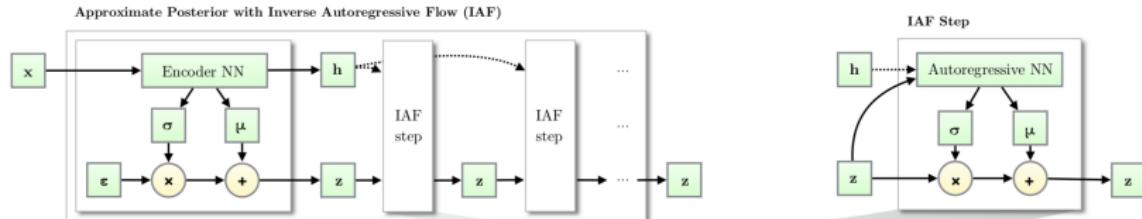
$$KL(p||\pi) = \mathbb{E}_{p(\mathbf{z})} \left[\log p(\mathbf{z}) - \log \left| \det \left(\frac{\partial g(\mathbf{z}, \boldsymbol{\theta})}{\partial \mathbf{z}} \right) \right| - \log \pi(g(\mathbf{z}, \boldsymbol{\theta})) \right]$$

- ▶ We don't need to think about computing the function $f(\mathbf{x}, \boldsymbol{\theta})$.
- ▶ Inverse autoregressive flow is a natural choice for using flows in VAE:

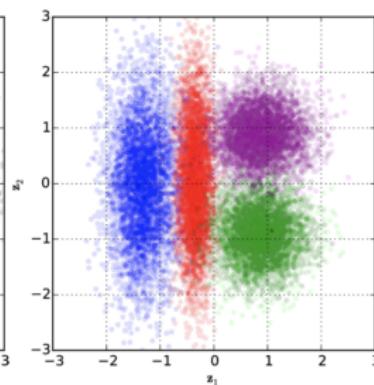
$$\mathbf{z} = \boldsymbol{\sigma}(\mathbf{x}) \odot \boldsymbol{\epsilon} + \boldsymbol{\mu}(\mathbf{x}), \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, 1); \quad \sim q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi}).$$

$$\mathbf{z}_k = \tilde{\boldsymbol{\sigma}}_k(\mathbf{z}_{k-1}) \odot \mathbf{z}_{k-1} + \tilde{\boldsymbol{\mu}}_k(\mathbf{z}_{k-1}), \quad k \geq 1; \quad \sim q_k(\mathbf{z}_k|\mathbf{x}, \boldsymbol{\phi}, \{\boldsymbol{\lambda}_j\}_{j=1}^k).$$

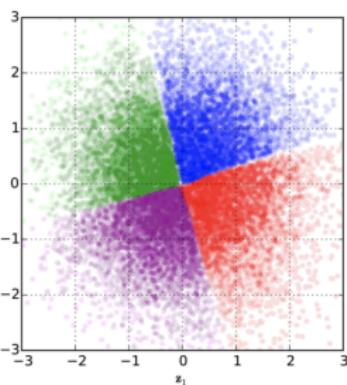
Inverse autoregressive flow (IAF)



(a) Prior distribution



(b) Posteriors in standard VAE



(c) Posteriors in VAE with IAF

Flows-based VAE prior vs posterior

Theorem

VAE with the flow-based prior for latent code \mathbf{z} is equivalent to using flow-based posterior for latent code ϵ .

Proof

$$\begin{aligned}\mathcal{L}(\phi, \theta, \lambda) &= \mathbb{E}_{q(\mathbf{z}|\mathbf{x}, \phi)} \log p(\mathbf{x}|\mathbf{z}, \theta) - \underbrace{KL(q(\mathbf{z}|\mathbf{x}, \phi)||p(\mathbf{z}|\lambda))}_{\text{flow-based prior}} \\ &= \mathbb{E}_{q(\mathbf{z}|\mathbf{x}, \phi)} \log p(\mathbf{x}|\mathbf{z}, \theta) - \underbrace{KL(q(\mathbf{z}|\mathbf{x}, \phi, \lambda)||p(\mathbf{z}))}_{\text{flow-based posterior}}\end{aligned}$$

(Here we use Flow KL duality theorem from Lecture 4)

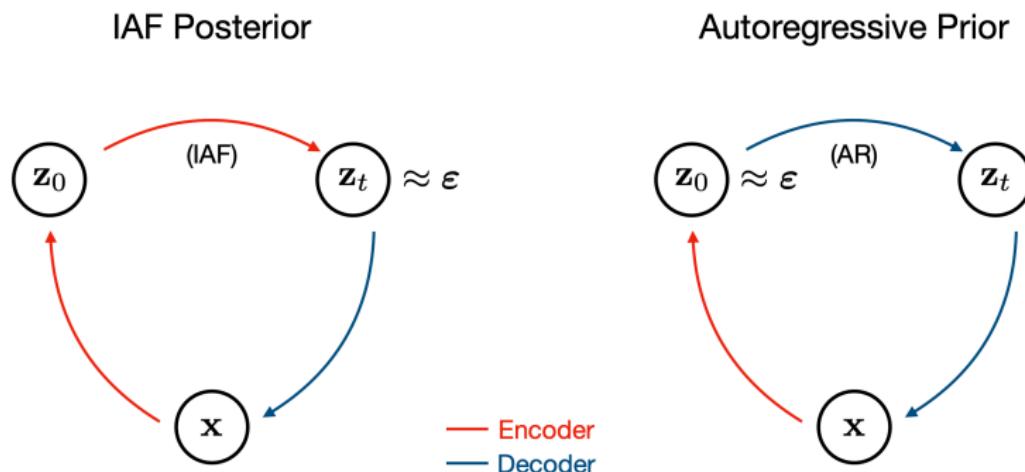
Flows in VAE posterior

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

Flows-based VAE prior vs posterior

- ▶ 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.



Recap of previous lecture

ELBO

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

- ▶ Normal variational distribution
 $q(\mathbf{z}|\mathbf{x}, \phi) = \mathcal{N}(\mathbf{z}|\mu_\phi(\mathbf{x}), \sigma_\phi^2(\mathbf{x}))$ is poor (e.g. has only one mode).
- ▶ Flows models convert a simple base distribution to a complex one using an invertible transformation with simple Jacobian.

Flow model in latent space

$$\log q_K(\mathbf{z}_K|\mathbf{x}, \phi_*) = \log q(\mathbf{z}_0|\mathbf{x}, \phi) - \sum_{k=1}^K \log \left| \det \left(\frac{\partial g_k(\mathbf{z}_{k-1}, \phi_k)}{\partial \mathbf{z}_{k-1}} \right) \right|.$$

Let's use $q_K(\mathbf{z}_K|\mathbf{x}, \phi_*)$, $\phi_* = \{\phi, \phi_1, \dots, \phi_K\}$ as a variational distribution. Here, ϕ – encoder parameters, $\{\phi_k\}_{k=1}^K$ – flow parameters.

Recap of previous lecture

Variational distribution

$$\log q_K(\mathbf{z}_K | \mathbf{x}, \phi_*) = \log q(\mathbf{z}_0 | \mathbf{x}, \phi) - \sum_{k=1}^K \log \left| \det \left(\frac{\partial g_k(\mathbf{z}_{k-1}, \phi_k)}{\partial \mathbf{z}_{k-1}} \right) \right|.$$

ELBO objective

$$\begin{aligned} \mathcal{L}(\phi, \theta) = & \mathbb{E}_{q(\mathbf{z}_0 | \mathbf{x}, \phi)} \left[\log p(\mathbf{x}, \mathbf{z}_K | \theta) - \log q(\mathbf{z}_0 | \mathbf{x}, \phi) + \right. \\ & \left. + \sum_{k=1}^K \log \left| \det \left(\frac{\partial g_k(\mathbf{z}_{k-1}, \phi_k)}{\partial \mathbf{z}_{k-1}} \right) \right| \right]. \end{aligned}$$

- ▶ Obtain samples \mathbf{z}_0 from the encoder.
- ▶ Apply flow model $\mathbf{z}_K = g(\mathbf{z}_0, \{\phi_k\}_{k=1}^K)$.
- ▶ Compute likelihood for \mathbf{z}_K using the decoder, base distribution for \mathbf{z}_0 and the Jacobian.
- ▶ We do not need an inverse flow function if we use flows in variational inference.

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

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

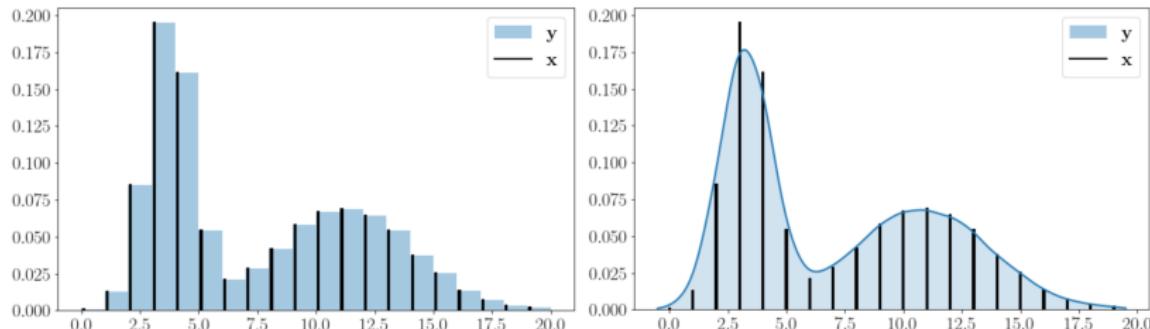
$$\log P(\mathbf{x}|\theta) \geq \int_{U[0,1]} \log p(\mathbf{x} + \mathbf{u}|\theta) d\mathbf{u}.$$

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

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

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} .

Dequantization

- ▶ Images are discrete data, pixels lie in the $[0, 255]$ integer domain (the model is $P(\mathbf{x}|\theta) = \text{Categorical}(\pi(\theta))$).
- ▶ Flow is a continuous model (it works with continuous data \mathbf{x}).

By fitting a continuous density model to discrete data, one can produce a degenerate solution with all probability mass on discrete values.

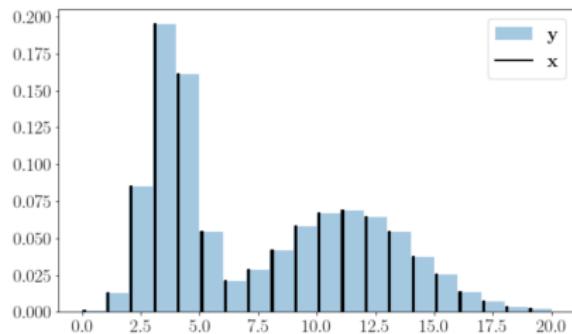
How to convert a discrete data distribution to a continuous one?

Uniform dequantization

$$\mathbf{x} \sim \text{Categorical}(\boldsymbol{\pi})$$

$$\mathbf{u} \sim U[0, 1]$$

$$\mathbf{y} = \mathbf{x} + \mathbf{u} \sim \text{Continuous}$$



Uniform dequantization

Statement

Fitting continuous model $p(\mathbf{y}|\theta)$ on uniformly dequantized data $\mathbf{y} = \mathbf{x} + \mathbf{u}$, $\mathbf{u} \sim U[0, 1]$ is equivalent to maximization of a lower bound on log-likelihood for a discrete model:

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

Thus, the maximisation of continuous model log-likelihood on \mathbf{y} can't lead to the a collapse onto the discrete data (the objective is bounded above by the discrete model log-likelihood).

Proof

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

Dequantization

- ▶ Images are discrete data, pixels lie in the $\{0, 255\}$ integer domain (the model is $P(\mathbf{x}|\theta) = \text{Categorical}(\pi(\theta))$).
- ▶ Flow is a continuous model (it works with continuous data \mathbf{x}).

By fitting a continuous density model to discrete data, one can produce a degenerate solution with all probability mass on discrete values.

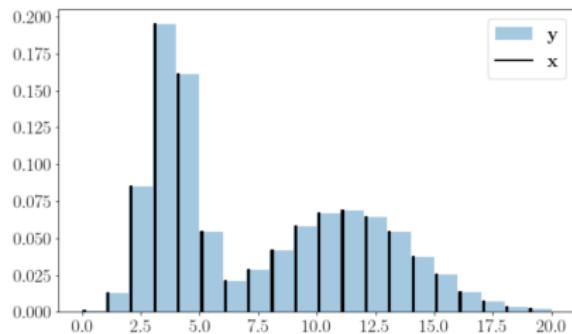
How to convert a discrete data distribution to a continuous one?

Uniform dequantization

$$\mathbf{x} \sim \text{Categorical}(\boldsymbol{\pi})$$

$$\mathbf{u} \sim U[0, 1]$$

$$\mathbf{y} = \mathbf{x} + \mathbf{u} \sim \text{Continuous}$$



Uniform dequantization

Statement

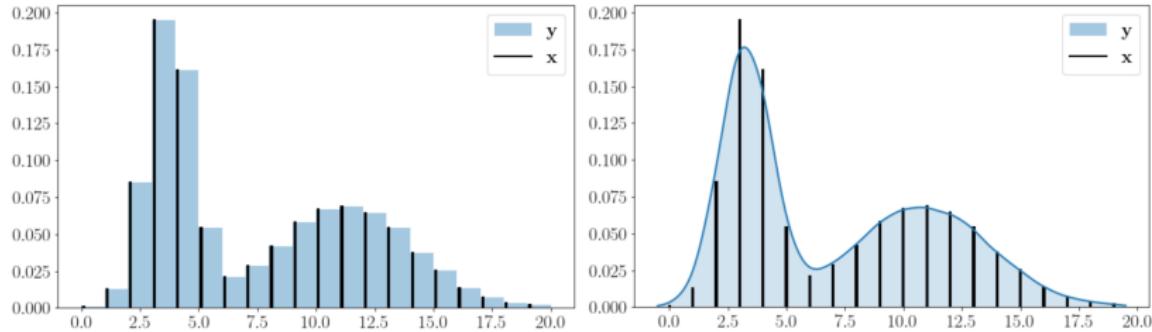
Fitting continuous model $p(\mathbf{y}|\theta)$ on uniformly dequantized data $\mathbf{y} = \mathbf{x} + \mathbf{u}$, $\mathbf{u} \sim U[0, 1]$ is equivalent to maximization of a lower bound on log-likelihood for a discrete model:

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

Proof

$$\begin{aligned}\mathbb{E}_\pi \log p(\mathbf{y}|\theta) &= \int \pi(\mathbf{y}) \log p(\mathbf{y}|\theta) d\mathbf{y} = \\ &= \sum \pi(\mathbf{x}) \int_{U[0,1]} \log p(\mathbf{x} + \mathbf{u}|\theta) d\mathbf{u} \leq \\ &\leq \sum \pi(\mathbf{x}) \log \int_{U[0,1]} p(\mathbf{x} + \mathbf{u}|\theta) d\mathbf{u} = \\ &= \sum \pi(\mathbf{x}) \log P(\mathbf{x}|\theta) = \mathbb{E}_\pi \log P(\mathbf{x}|\theta).\end{aligned}$$

Variational dequantization



- ▶ $p(y|\theta)$ assign uniform density to unit hypercubes $x + U[0, 1]$ (left fig).
- ▶ Neural network density models are smooth function approximators (right fig).
- ▶ Smooth dequantization is more natural.

How to perform the smooth dequantization?

Flow++

Variational dequantization

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

Variational lower bound

$$\begin{aligned}\log P(\mathbf{x}|\theta) &= \left[\log \int q(\mathbf{u}|\mathbf{x}) \frac{p(\mathbf{x} + \mathbf{u}|\theta)}{q(\mathbf{u}|\mathbf{x})} d\mathbf{u} \right] \geq \\ &\geq \int q(\mathbf{u}|\mathbf{x}) \log \frac{p(\mathbf{x} + \mathbf{u}|\theta)}{q(\mathbf{u}|\mathbf{x})} d\mathbf{u} = \mathcal{L}(q, \theta).\end{aligned}$$

Uniform dequantization bound

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

Flow++

Variational lower bound

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

Let $\mathbf{u} = h(\epsilon, \phi)$ is a flow model with base distribution $\epsilon \sim p(\epsilon) = \mathcal{N}(0, \mathbf{I})$:

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

Then

$$\log P(\mathbf{x}|\theta) \geq \mathcal{L}(\phi, \theta) = \int p(\epsilon) \log \left(\frac{p(\mathbf{x} + h(\epsilon, \phi)|\theta)}{p(\epsilon) \cdot \left| \det \frac{\partial h(\epsilon, \phi)}{\partial \epsilon} \right|^{-1}} \right) d\epsilon.$$

Variational lower

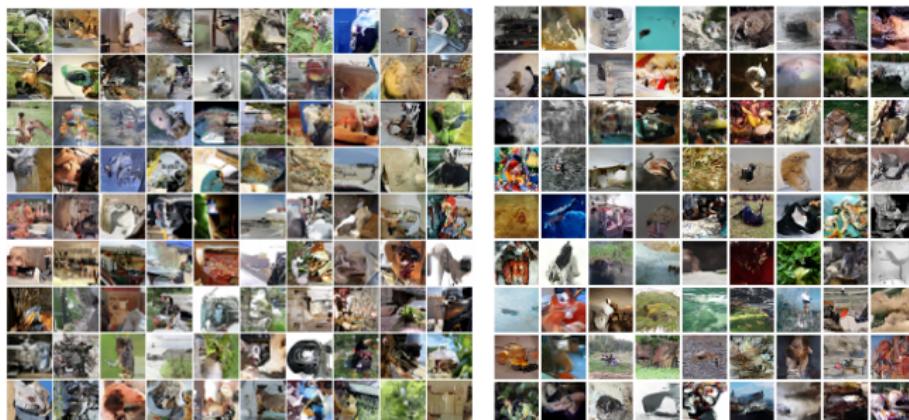
$$\log P(\mathbf{x}|\theta) \geq \int p(\epsilon) \log \left(\frac{p(\mathbf{x} + h(\epsilon, \phi))}{p(\epsilon) \cdot \left| \det \frac{\partial h(\epsilon, \phi)}{\partial \epsilon} \right|^{-1}} \right) d\epsilon.$$

- ▶ If $p(\mathbf{x} + \mathbf{u}|\theta)$ is also a flow model, it is straightforward to calculate stochastic gradient of this ELBO.
- ▶ Uniform dequantization is a special case of variational dequantization ($q(\mathbf{u}|\mathbf{x}) = U[0, 1]$). The gap between $\log P(\mathbf{x}|\theta)$ and the derived ELBO is $KL(q(\mathbf{u}|\mathbf{x})||p(\mathbf{u}|\mathbf{x}))$.
- ▶ In the case of uniform dequantization the model unnaturally places uniform density over each hypercube $\mathbf{x} + U[0, 1]$ due to inexpressive distribution q .

Flow++

Table 1. Unconditional image modeling results in bits/dim

Model family	Model	CIFAR10	ImageNet 32x32	ImageNet 64x64
Non-autoregressive	RealNVP (Dinh et al., 2016)	3.49	4.28	—
	Glow (Kingma & Dhariwal, 2018)	3.35	4.09	3.81
	IAF-VAE (Kingma et al., 2016)	3.11	—	—
	Flow++ (ours)	3.08	3.86	3.69
Autoregressive	Multiscale PixelCNN (Reed et al., 2017)	—	3.95	3.70
	PixelCNN (van den Oord et al., 2016b)	3.14	—	—
	PixelRNN (van den Oord et al., 2016b)	3.00	3.86	3.63
	Gated PixelCNN (van den Oord et al., 2016c)	3.03	3.83	3.57
	PixelCNN++ (Salimans et al., 2017)	2.92	—	—
	Image Transformer (Parmar et al., 2018)	2.90	3.77	—
	PixelSNAIL (Chen et al., 2017)	2.85	3.80	3.52



(a) PixelCNN

(b) Flow++

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

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

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

- ▶ $\pi(\mathbf{x}|\mathbf{v}, \mathbf{w}) = \text{Sim}(\mathbf{v}, \mathbf{w})$ – true world simulator;
- ▶ \mathbf{v} – conditionally independent factors: $\pi(\mathbf{v}|\mathbf{x}) = \prod_{j=1}^d \pi(v_j|\mathbf{x})$;
- ▶ \mathbf{w} – conditionally dependent factors.

Goal

Construct an unsupervised deep generative model

$$p(\mathbf{x}|\mathbf{z}, \theta) \approx \pi(\mathbf{x}|\mathbf{v}, \mathbf{w}).$$

- ▶ 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

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

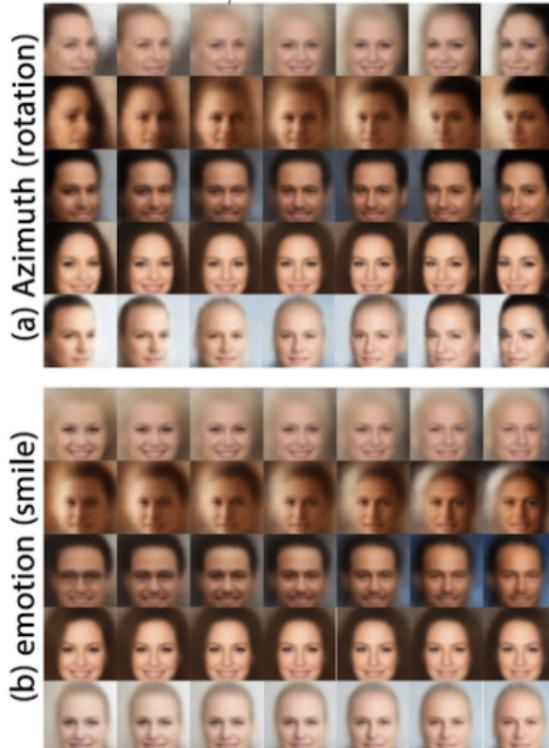
Hypothesis

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

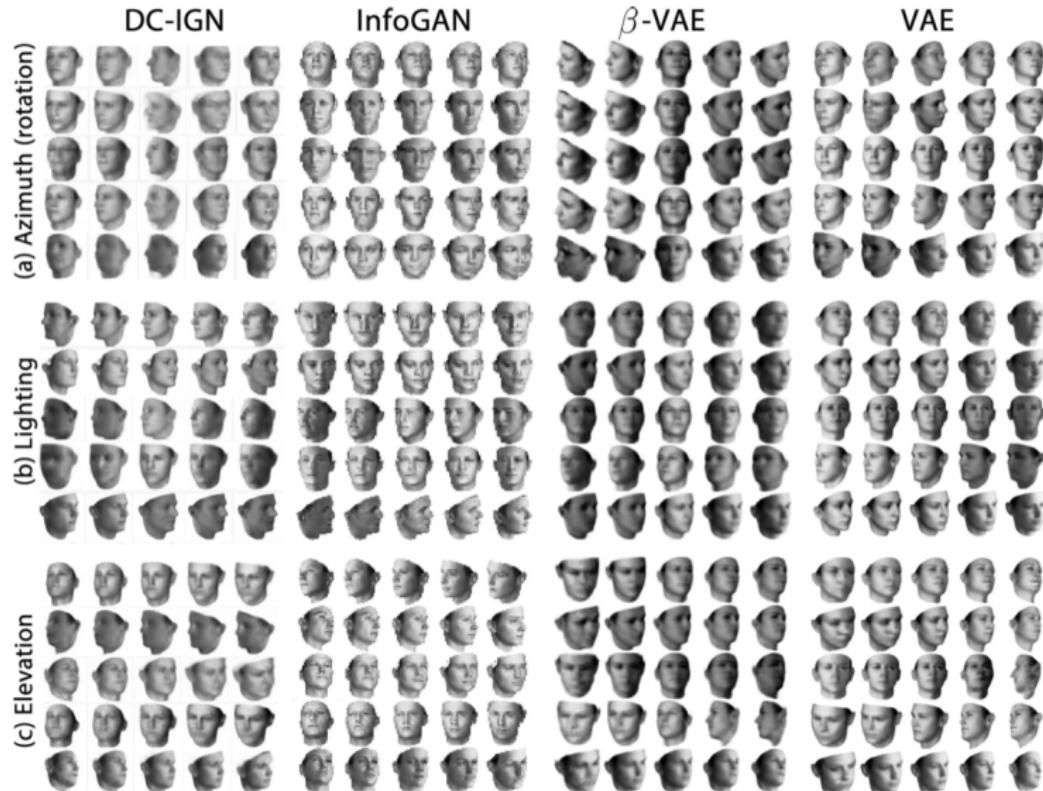
β -VAE



VAE

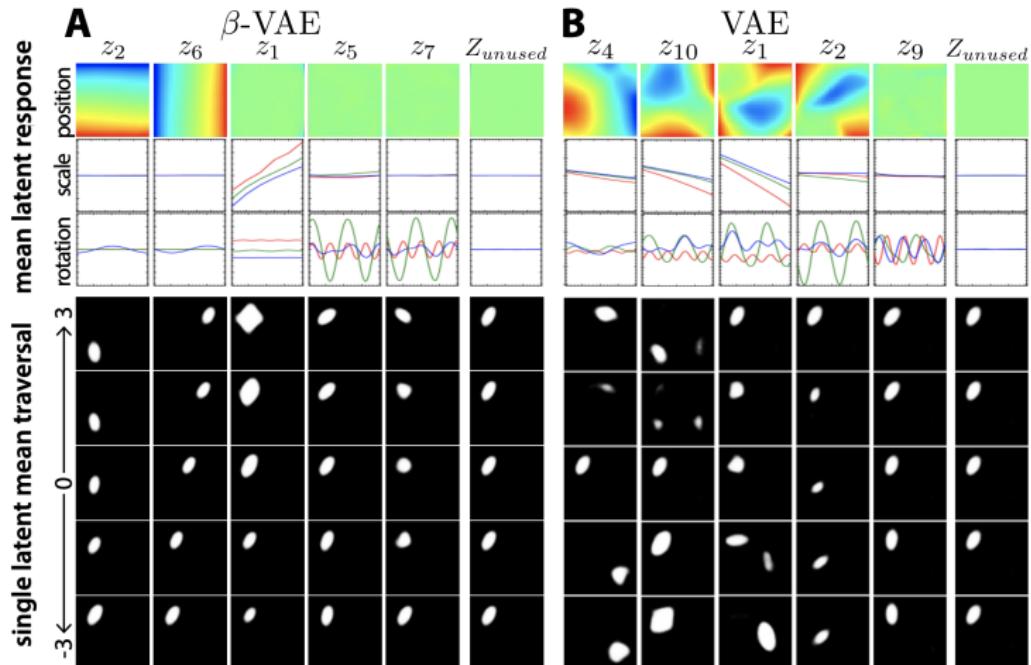


β -VAE



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.

Summary

- ▶ Uniform dequantization is the simplest form of dequantization.
- ▶ Variational dequantization is a more natural type that was proposed in Flow++ model.
- ▶ Disentanglement learning tries to make latent components more informative.
- ▶ β -VAE makes the latent components more independent, but the reconstructions get poorer.