

# Deep Generative Models

## Lecture 6

Roman Isachenko

Moscow Institute of Physics and Technology

2023, Autumn

## Recap of previous lecture

Let split  $\mathbf{x}$  and  $\mathbf{z}$  in two parts:

$$\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2] = [\mathbf{x}_{1:d}, \mathbf{x}_{d+1:m}]; \quad \mathbf{z} = [\mathbf{z}_1, \mathbf{z}_2] = [\mathbf{z}_{1:d}, \mathbf{z}_{d+1:m}].$$

## Coupling layer

$$\begin{cases} \mathbf{x}_1 = \mathbf{z}_1; \\ \mathbf{x}_2 = \mathbf{z}_2 \odot \sigma_{\theta}(\mathbf{z}_1) + \mu_{\theta}(\mathbf{z}_1). \end{cases} \quad \begin{cases} \mathbf{z}_1 = \mathbf{x}_1; \\ \mathbf{z}_2 = (\mathbf{x}_2 - \mu_{\theta}(\mathbf{x}_1)) \odot \frac{1}{\sigma_{\theta}(\mathbf{x}_1)}. \end{cases}$$

Estimating the density takes 1 pass, sampling takes 1 pass!

## Jacobian

$$\det \left( \frac{\partial \mathbf{z}}{\partial \mathbf{x}} \right) = \det \begin{pmatrix} \mathbf{I}_d & 0_{d \times m-d} \\ \frac{\partial \mathbf{z}_2}{\partial \mathbf{x}_1} & \frac{\partial \mathbf{z}_2}{\partial \mathbf{x}_2} \end{pmatrix} = \prod_{j=1}^{m-d} \frac{1}{\sigma_j(\mathbf{x}_1)}.$$

Coupling layer is a special case of autoregressive NF.

## Recap of previous lecture

	VAE	NF
<b>Objective</b>	ELBO $\mathcal{L}$	Forward KL/MLE
<b>Encoder</b>	stochastic $\mathbf{z} \sim q(\mathbf{z} \mathbf{x}, \phi)$	deterministic $\mathbf{z} = f_\theta(\mathbf{x})$ $q(\mathbf{z} \mathbf{x}, \theta) = \delta(\mathbf{z} - f_\theta(\mathbf{x}))$
<b>Decoder</b>	stochastic $\mathbf{x} \sim p(\mathbf{x} \mathbf{z}, \theta)$	deterministic $\mathbf{x} = g_\theta(\mathbf{z})$ $p(\mathbf{x} \mathbf{z}, \theta) = \delta(\mathbf{x} - g_\theta(\mathbf{z}))$
<b>Parameters</b>	$\phi, \theta$	$\theta \equiv \phi$

### Theorem

MLE for normalizing flow is equivalent to maximization of ELBO for VAE model with deterministic encoder and decoder:

$$p(\mathbf{x}|\mathbf{z}, \theta) = \delta(\mathbf{x} - f^{-1}(\mathbf{z}, \theta)) = \delta(\mathbf{x} - g_\theta(\mathbf{z}));$$

$$q(\mathbf{z}|\mathbf{x}, \theta) = p(\mathbf{z}|\mathbf{x}, \theta) = \delta(\mathbf{z} - f_\theta(\mathbf{x})).$$

## Recap of previous lecture

Let our data  $\mathbf{y}$  comes from discrete distribution  $\Pi(\mathbf{y})$ .

### Discrete model

- ▶ Use **discrete** model (e.x.  $P(\mathbf{y}|\theta) = \text{Cat}(\pi(\theta))$ ).
- ▶ Minimize any suitable divergence measure  $D(\Pi, P)$ .

### Continuous model

Use **continuous** model (e.x.  $p(\mathbf{x}|\theta) = \mathcal{N}(\mu_\theta(\mathbf{x}), \sigma_\theta^2(\mathbf{x}))$ ), but

- ▶ **discretize** model (make the model outputs discrete): transform  $p(\mathbf{x}|\theta)$  to  $P(\mathbf{y}|\theta)$ ;
- ▶ **dequantize** data (make the data continuous): transform  $\Pi(\mathbf{y})$  to  $\pi(\mathbf{x})$ .

### Model discretization through CDF

$$F(\mathbf{x}|\theta) = \int_{-\infty}^{\mathbf{x}} p(\mathbf{x}'|\theta) d\mathbf{x}'; \quad P(\mathbf{y}|\theta) = F(\mathbf{y} + 0.5|\theta) - F(\mathbf{y} - 0.5|\theta)$$

## Recap of previous lecture

### Uniform dequantization bound

Let dequantize discrete distribution  $\Pi(\mathbf{y})$  to continuous distribution  $\pi(\mathbf{x})$  in the following way:  $\mathbf{x} = \mathbf{y} + \mathbf{u}$ , where  $\mathbf{u} \sim U[0, 1]$ .

### Theorem

Fitting continuous model  $p(\mathbf{x}|\theta)$  on uniformly dequantized data is equivalent to maximization of a lower bound on log-likelihood for a discrete model:

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

### Variational dequantization bound

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

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

# Outline

1. ELBO surgery
2. Learnable VAE prior
3. Discrete VAE latent representations  
Vector quantization

# Outline

1. ELBO surgery
2. Learnable VAE prior
3. Discrete VAE latent representations  
Vector quantization

# ELBO surgery

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

## Theorem

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

- ▶  $q_{\text{agg}}(\mathbf{z}) = \frac{1}{n} \sum_{i=1}^n q(\mathbf{z}|\mathbf{x}_i)$  – **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  $q_{\text{agg}}(\mathbf{z})$  towards the prior  $p(\mathbf{z})$ .
- ▶ **Second term** reduces the amount of information about  $\mathbf{x}$  stored in  $\mathbf{z}$ .

# ELBO surgery

## Theorem

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

## Proof

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^n KL(q(\mathbf{z}|\mathbf{x}_i) || p(\mathbf{z})) &= \frac{1}{n} \sum_{i=1}^n \int q(\mathbf{z}|\mathbf{x}_i) \log \frac{q(\mathbf{z}|\mathbf{x}_i)}{p(\mathbf{z})} d\mathbf{z} = \\ &= \frac{1}{n} \sum_{i=1}^n \int q(\mathbf{z}|\mathbf{x}_i) \log \frac{q_{\text{agg}}(\mathbf{z}) q(\mathbf{z}|\mathbf{x}_i)}{p(\mathbf{z}) q_{\text{agg}}(\mathbf{z})} d\mathbf{z} = \int \frac{1}{n} \sum_{i=1}^n q(\mathbf{z}|\mathbf{x}_i) \log \frac{q_{\text{agg}}(\mathbf{z})}{p(\mathbf{z})} d\mathbf{z} + \\ &+ \frac{1}{n} \sum_{i=1}^n \int q(\mathbf{z}|\mathbf{x}_i) \log \frac{q(\mathbf{z}|\mathbf{x}_i)}{q_{\text{agg}}(\mathbf{z})} d\mathbf{z} = KL(q_{\text{agg}}(\mathbf{z}) || p(\mathbf{z})) + \frac{1}{n} \sum_{i=1}^n KL(q(\mathbf{z}|\mathbf{x}_i) || q_{\text{agg}}(\mathbf{z})) \end{aligned}$$

Without proof:

$$\mathbb{I}_q[\mathbf{x}, \mathbf{z}] = \frac{1}{n} \sum_{i=1}^n KL(q(\mathbf{z}|\mathbf{x}_i) || q_{\text{agg}}(\mathbf{z})) \in [0, \log n].$$

# ELBO surgery

## ELBO revisiting

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

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

## Optimal VAE prior

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

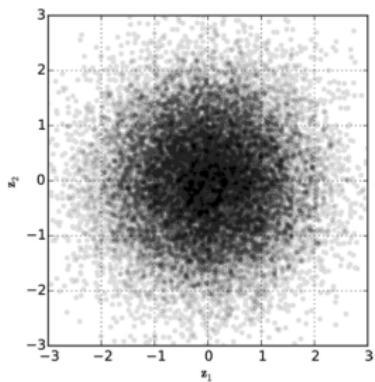
The optimal prior  $p(\mathbf{z})$  is the aggregated posterior  $q_{\text{agg}}(\mathbf{z})$ !

# Variational posterior

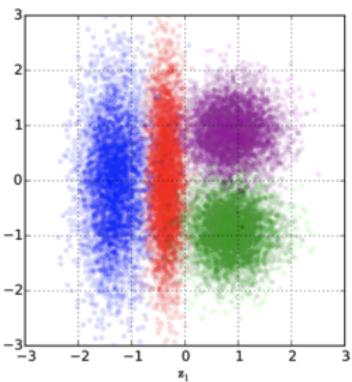
## ELBO decomposition

$$\log p(\mathbf{x}|\theta) = \mathcal{L}(q, \theta) + KL(q(\mathbf{z}|\mathbf{x}, \phi)||p(\mathbf{z}|\mathbf{x}, \theta)).$$

- ▶  $q(\mathbf{z}|\mathbf{x}, \phi) = \mathcal{N}(\mathbf{z}|\mu_\phi(\mathbf{x}), \sigma_\phi^2(\mathbf{x}))$  is a unimodal distribution (not expressive enough).
- ▶ The optimal prior  $p(\mathbf{z})$  is the aggregated posterior  $q_{\text{agg}}(\mathbf{z})$ .



(a) Prior distribution



(b) Posteriors in standard VAE

# Outline

1. ELBO surgery
2. Learnable VAE prior
3. Discrete VAE latent representations  
Vector quantization

## VAE limitations

- ▶ Poor generative distribution (decoder)

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

- ▶ Loose lower bound

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

- ▶ **Poor prior distribution**

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

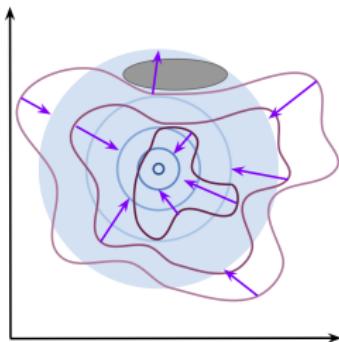
- ▶ Poor variational posterior distribution (encoder)

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

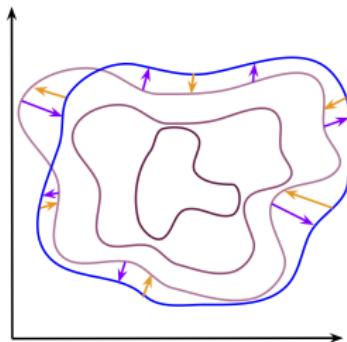
# Optimal VAE prior

- ▶ Standard Gaussian  $p(\mathbf{z}) = \mathcal{N}(0, I) \Rightarrow$  over-regularization;
- ▶  $p(\mathbf{z}) = q_{\text{agg}}(\mathbf{z}) = \frac{1}{n} \sum_{i=1}^n q(\mathbf{z}|\mathbf{x}_i) \Rightarrow$  overfitting and highly expensive.

Non learnable prior  $p(\mathbf{z})$



Learnable prior  $p(\mathbf{z}|\boldsymbol{\lambda})$



ELBO revisiting

$$\frac{1}{n} \sum_{i=1}^n \mathcal{L}_i(q, \theta) = \text{RL} - \text{MI} - KL(q_{\text{agg}}(\mathbf{z}) || p(\mathbf{z}|\boldsymbol{\lambda}))$$

It is Forward KL with respect to  $p(\mathbf{z}|\boldsymbol{\lambda})$ .

## NF-based VAE prior

NF model in latent space

$$\log p(\mathbf{z}|\boldsymbol{\lambda}) = \log p(\mathbf{z}^*) + \log \left| \det \left( \frac{d\mathbf{z}^*}{d\mathbf{z}} \right) \right| = \log p(f(\mathbf{z}, \boldsymbol{\lambda})) + \log |\det(\mathbf{J}_f)|$$

$$\mathbf{z} = g_{\boldsymbol{\lambda}}(\mathbf{z}^*) = f_{\boldsymbol{\lambda}}^{-1}(\mathbf{z}^*)$$

- ▶ RealNVP with coupling layers.
- ▶ Autoregressive NF (fast  $f_{\boldsymbol{\lambda}}(\mathbf{z})$ , slow  $g_{\boldsymbol{\lambda}}(\mathbf{z}^*)$ ).

ELBO with NF-based VAE prior

$$\begin{aligned}\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\theta}) &= \mathbb{E}_{q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi})} [\log p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) + \log p(\mathbf{z}|\boldsymbol{\lambda}) - \log q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi})] \\ &= \mathbb{E}_{q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi})} \left[ \log p(\mathbf{x}|\mathbf{z}, \boldsymbol{\theta}) + \underbrace{\left( \log p(f_{\boldsymbol{\lambda}}(\mathbf{z})) + \log |\det(\mathbf{J}_f)| \right)}_{\text{NF-based prior}} - \log q(\mathbf{z}|\mathbf{x}, \boldsymbol{\phi}) \right]\end{aligned}$$

# Outline

1. ELBO surgery
2. Learnable VAE prior
3. Discrete VAE latent representations  
Vector quantization

# Discrete VAE latents

## Motivation

- ▶ Previous VAE models had **continuous** latent variables  $\mathbf{z}$ .
- ▶ **Discrete** representations  $\mathbf{z}$  are potentially a more natural fit for many of the modalities.
- ▶ Powerful autoregressive models (like PixelCNN) have been developed for modelling distributions over discrete variables.
- ▶ All cool transformer-like models work with discrete tokens.

## ELBO

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

- ▶ Reparametrization trick to get unbiased gradients.
- ▶ Normal assumptions for  $q(\mathbf{z}|\mathbf{x}, \phi)$  and  $p(\mathbf{z})$  to compute KL analytically.

# Discrete VAE latents

## Assumptions

- ▶ Define dictionary (word book) space  $\{\mathbf{e}_k\}_{k=1}^K$ , where  $\mathbf{e}_k \in \mathbb{R}^C$ ,  $K$  is the size of the dictionary.
- ▶ Let  $c \sim \text{Categorical}(\boldsymbol{\pi})$ , where
$$\boldsymbol{\pi} = (\pi_1, \dots, \pi_K), \quad \pi_k = P(c = k), \quad \sum_{k=1}^K \pi_k = 1.$$
- ▶ Let VAE model has discrete latent representation  $c$  with prior  $p(c) = \text{Uniform}\{1, \dots, K\}$ .

## How it should work?

- ▶ Our variational posterior  $q(c|\mathbf{x}, \phi) = \text{Categorical}(\boldsymbol{\pi}_\phi(\mathbf{x}))$  (encoder) outputs discrete probabilities vector.
- ▶ We sample  $c^*$  from  $q(c|\mathbf{x}, \phi)$  (reparametrization trick analogue).
- ▶ Our generative distribution  $p(\mathbf{x}|\mathbf{e}_{c^*}, \theta)$  (decoder).

# Discrete VAE latents

## ELBO

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

### KL term

$$\begin{aligned} KL(q(c|\mathbf{x}, \phi)||p(c)) &= \sum_{k=1}^K q(k|\mathbf{x}, \phi) \log \frac{q(k|\mathbf{x}, \phi)}{p(k)} = \\ &= \sum_{k=1}^K q(k|\mathbf{x}, \phi) \log q(k|\mathbf{x}, \phi) - \sum_{k=1}^K q(k|\mathbf{x}, \phi) \log p(k) = \\ &= -H(q(c|\mathbf{x}, \phi)) + \log K. \end{aligned}$$

- ▶ Is it possible to make reparametrization trick? (we sample from discrete distribution now!).
- ▶ Entropy term should be estimated.

# Outline

1. ELBO surgery
2. Learnable VAE prior
3. Discrete VAE latent representations  
Vector quantization

# Vector quantization

## Quantized representation

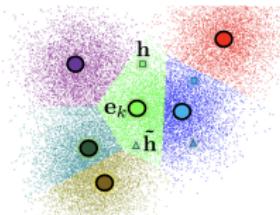
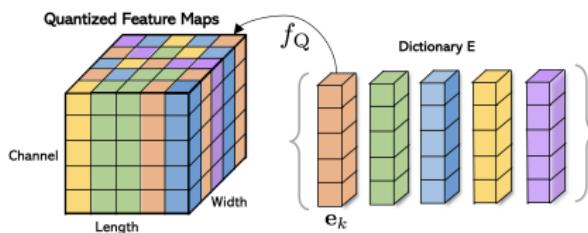
$\mathbf{z}_q \in \mathbb{R}^C$  for  $\mathbf{z} \in \mathbb{R}^C$  is defined by a nearest neighbor look-up using the shared dictionary space

$$\mathbf{z}_q = \mathbf{e}_{k^*}, \quad \text{where } k^* = \arg \min_k \|\mathbf{z} - \mathbf{e}_k\|.$$

- ▶ Let our encoder outputs continuous representation  $\mathbf{z}$ .
- ▶ Quantization will give us the discrete distribution  $q(c|x, \phi)$ .

## Quantization procedure

If we have tensor with the spatial dimensions we apply the quantization for each of  $W \times H$  locations.



## Vector Quantized VAE (VQ-VAE)

Let VAE latent variable  $\mathbf{c} \in \{1, \dots, K\}^{W \times H}$  is the discrete with spatial-independent variational posterior and prior distributions

$$q(\mathbf{c}|\mathbf{x}, \phi) = \prod_{i=1}^W \prod_{j=1}^H q(c_{ij}|\mathbf{x}, \phi); \quad p(\mathbf{c}) = \prod_{i=1}^W \prod_{j=1}^H \text{Uniform}\{1, \dots, K\}.$$

Let  $\mathbf{z}_e = \text{NN}_{e, \phi}(\mathbf{x}) \in \mathbb{R}^{W \times H \times C}$  is the encoder output.

### Deterministic variational posterior

$$q(c_{ij} = k^* | \mathbf{x}, \phi) = \begin{cases} 1, & \text{for } k^* = \arg \min_k \|[\mathbf{z}_e]_{ij} - \mathbf{e}_k\|; \\ 0, & \text{otherwise.} \end{cases}$$

$KL(q(c|\mathbf{x}, \phi)||p(c))$  term in ELBO is constant, entropy of the posterior is zero.

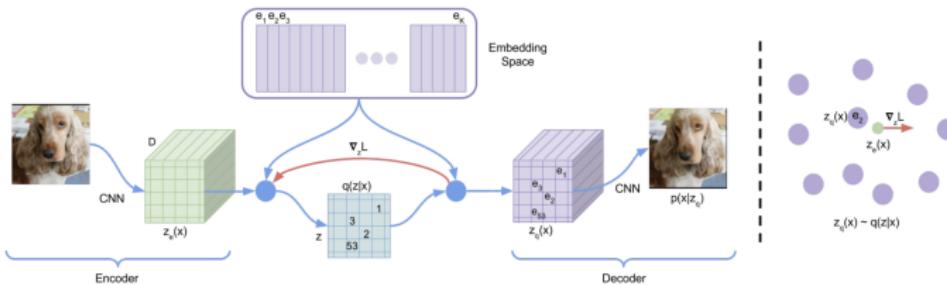
$$KL(q(c|\mathbf{x}, \phi)||p(c)) = -H(q(c|\mathbf{x}, \phi)) + \log K = \log K.$$

# Vector Quantized VAE (VQ-VAE)

## ELBO

$$\mathcal{L}(\phi, \theta) = \mathbb{E}_{q(c|x, \phi)} \log p(x|e_c, \theta) - \log K = \log p(x|z_q, \theta) - \log K,$$

where  $z_q = e_{k^*}$ ,  $k^* = \arg \min_k \|z_e - e_k\|$ .



**Problem:**  $\arg \min$  is not differentiable.

**Straight-through gradient estimation**

$$\frac{\partial \log p(x|z_q, \theta)}{\partial \phi} = \frac{\partial \log p(x|z_q, \theta)}{\partial z_q} \cdot \frac{\partial z_q}{\partial \phi} \approx \frac{\partial \log p(x|z_q, \theta)}{\partial z_q} \cdot \frac{\partial z_e}{\partial \phi}$$

# Vector Quantized VAE-2 (VQ-VAE-2)

Samples 1024x1024



Samples diversity



VQ-VAE (Proposed)

BigGAN deep

Razavi A., Oord A., Vinyals O. Generating Diverse High-Fidelity Images with VQ-VAE-2, 2019

## Summary

- ▶ The ELBO surgery reveals insights about a prior distribution in VAE. The optimal prior is the aggregated posterior.
- ▶ We could use NF-based prior in VAE (even autoregressive).
- ▶ Discrete VAE latents is a natural idea, but we have to avoid non-differentiable sampling operation.
- ▶ Vector Quantization is the way to create VAE with discrete latent space and deterministic variational posterior.
- ▶ Straight-through gradient ignores quantize operation in backprop.