VQ-VAE

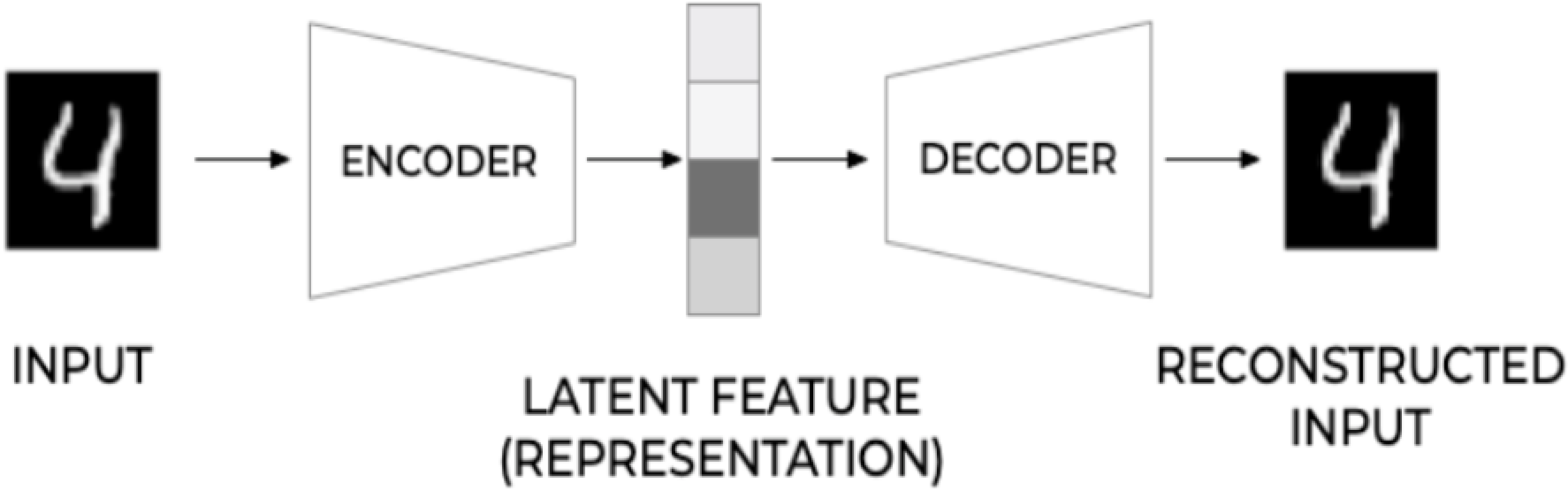
赵洲

浙江大学计算机学院

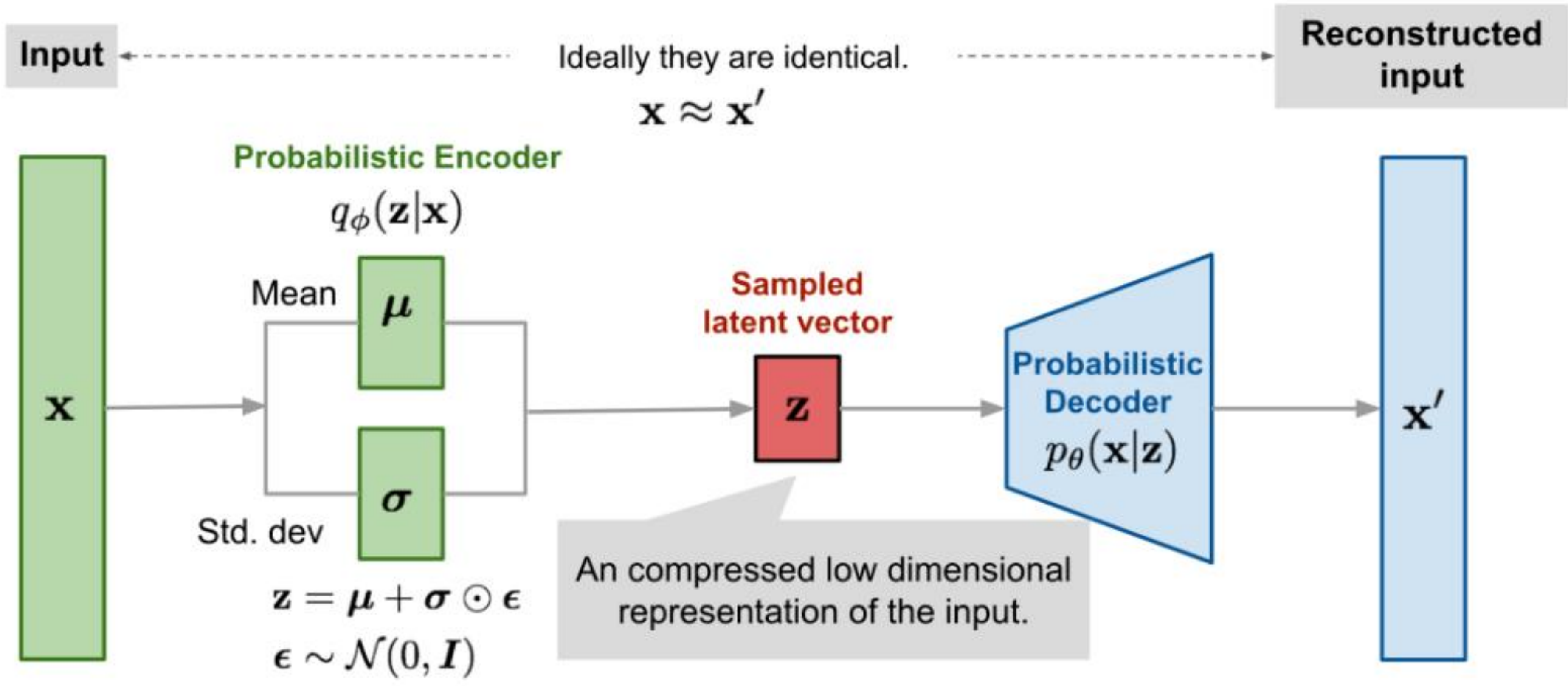
内容

- VAE复习
- VQ-VAE

AE


ENCODER
DECODER
INPUT
LATENT FEATURE
(REPRESENTATION)
RECONSTRUCTED
INPUT

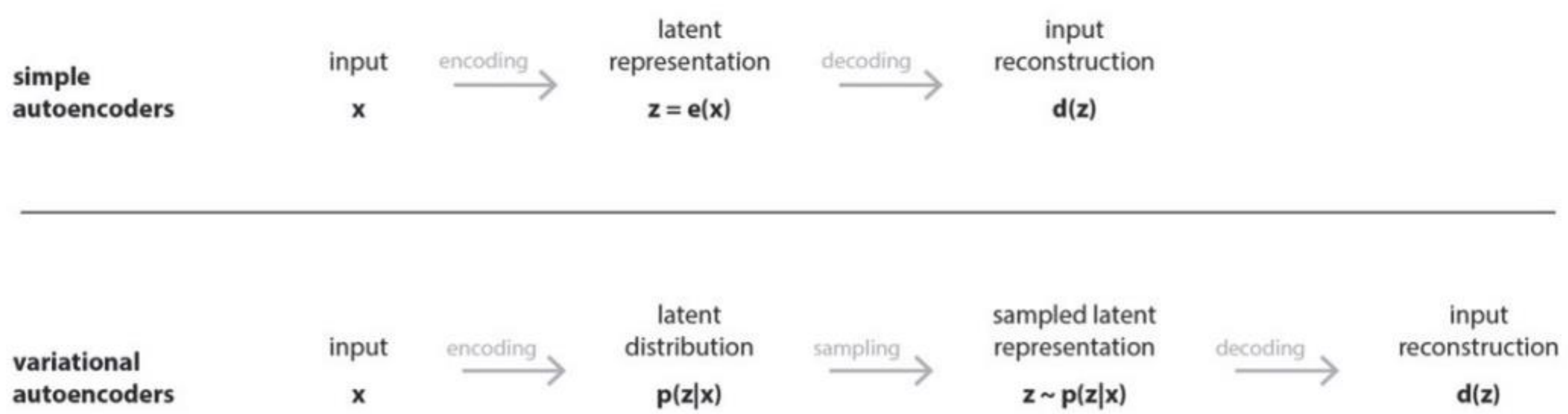
VAE



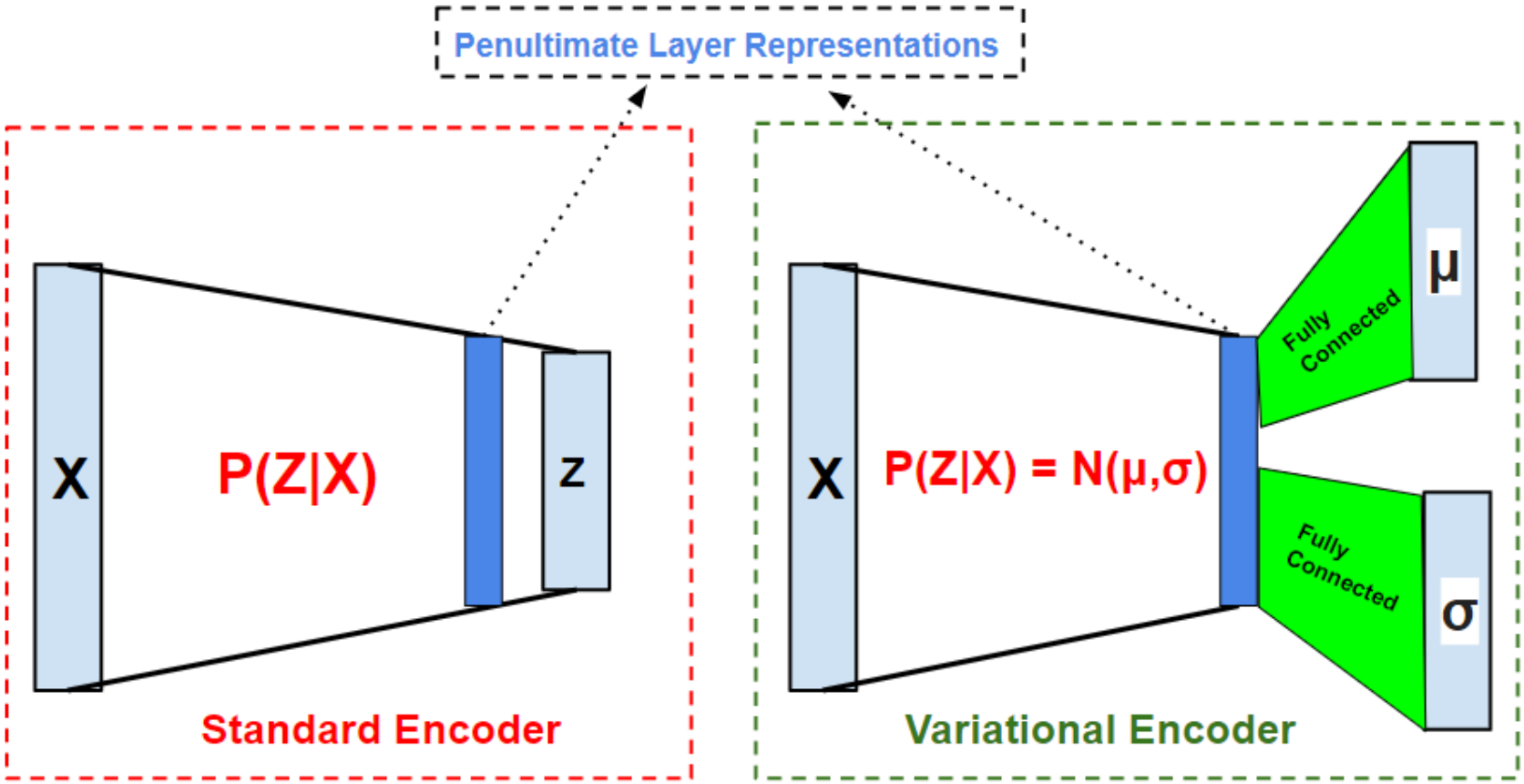
Input
Ideally they are identical.
$\mathbf{x} \approx \mathbf{x}'$
Reconstructed input
Probabilistic Encoder
$q_\phi(\mathbf{z}|\mathbf{x})$
$\mathbf{x}$
Mean
$\boldsymbol{\mu}$
Std. dev
$\boldsymbol{\sigma}$
$\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$
$\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I})$
Sampled latent vector
$\mathbf{z}$
An compressed low dimensional representation of the input.
Probabilistic Decoder
$p_\theta(\mathbf{x}|\mathbf{z})$
$\mathbf{x}'$

从AE到VAE

- 在AE的基础上，显性对z的分布p(z)进行建模，使得自编码器成为一个合格的生成模型（VAE）。

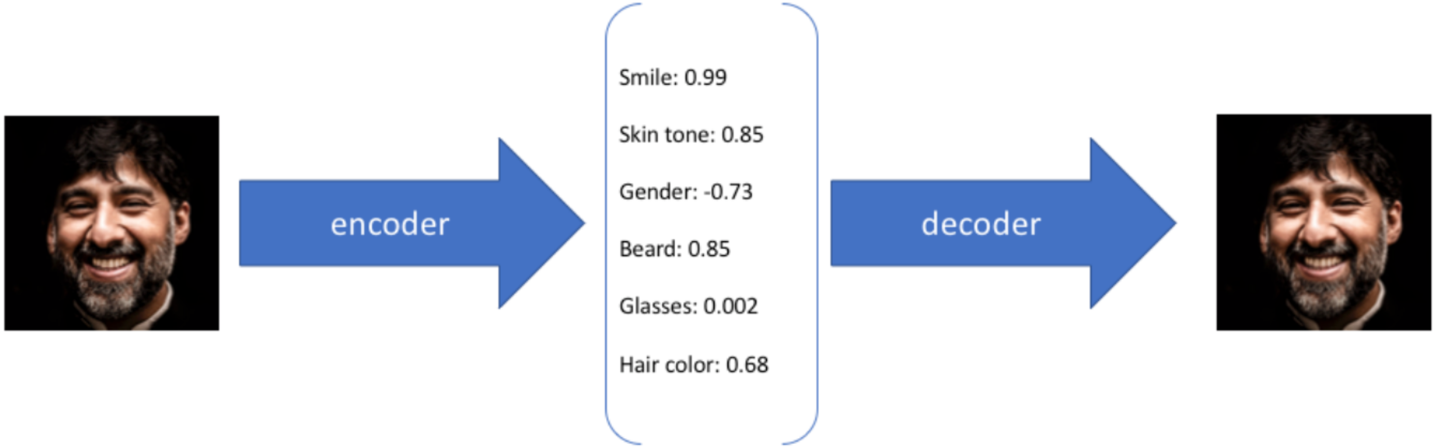


AE v.s. VAE编码

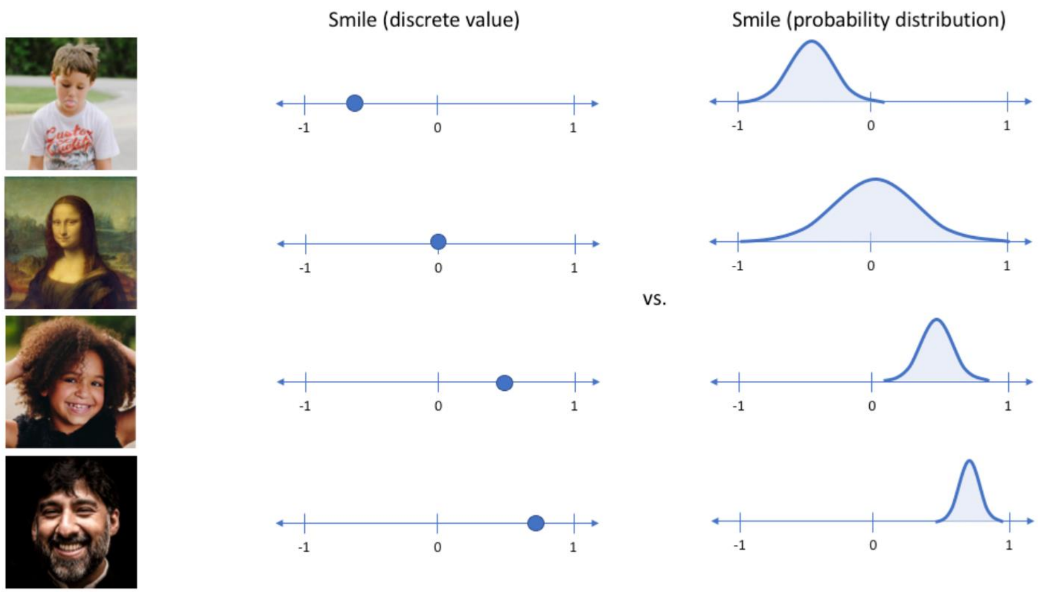


Penultimate Layer Representations
X
P(Z|X)
z
Standard Encoder
X
P(Z|X) = N(μ,σ)
Fully Connected
μ
Fully Connected
σ
Variational Encoder

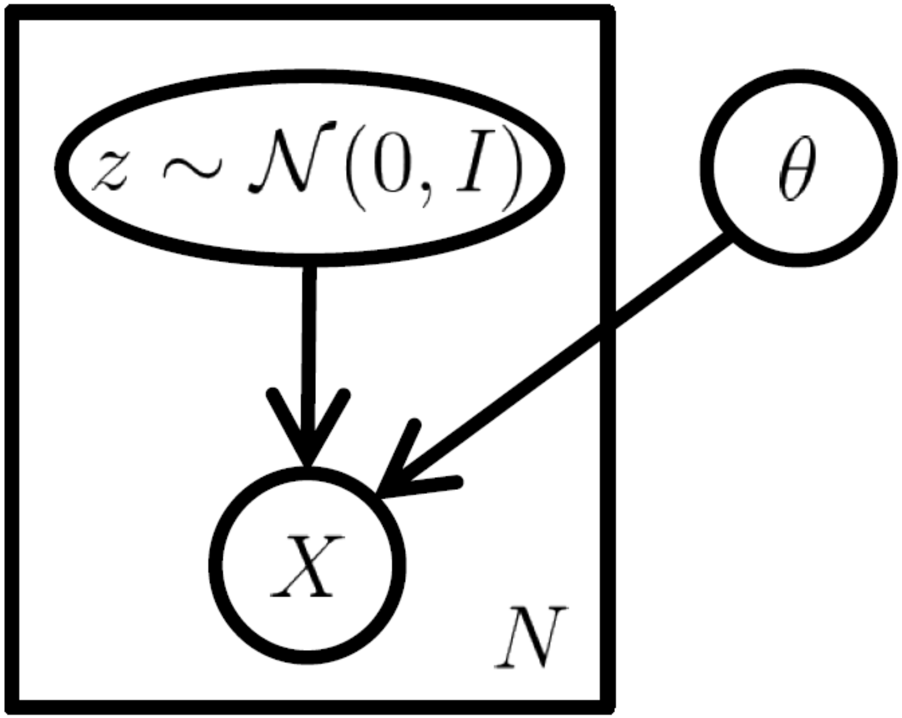
例子解释



Latent attributes

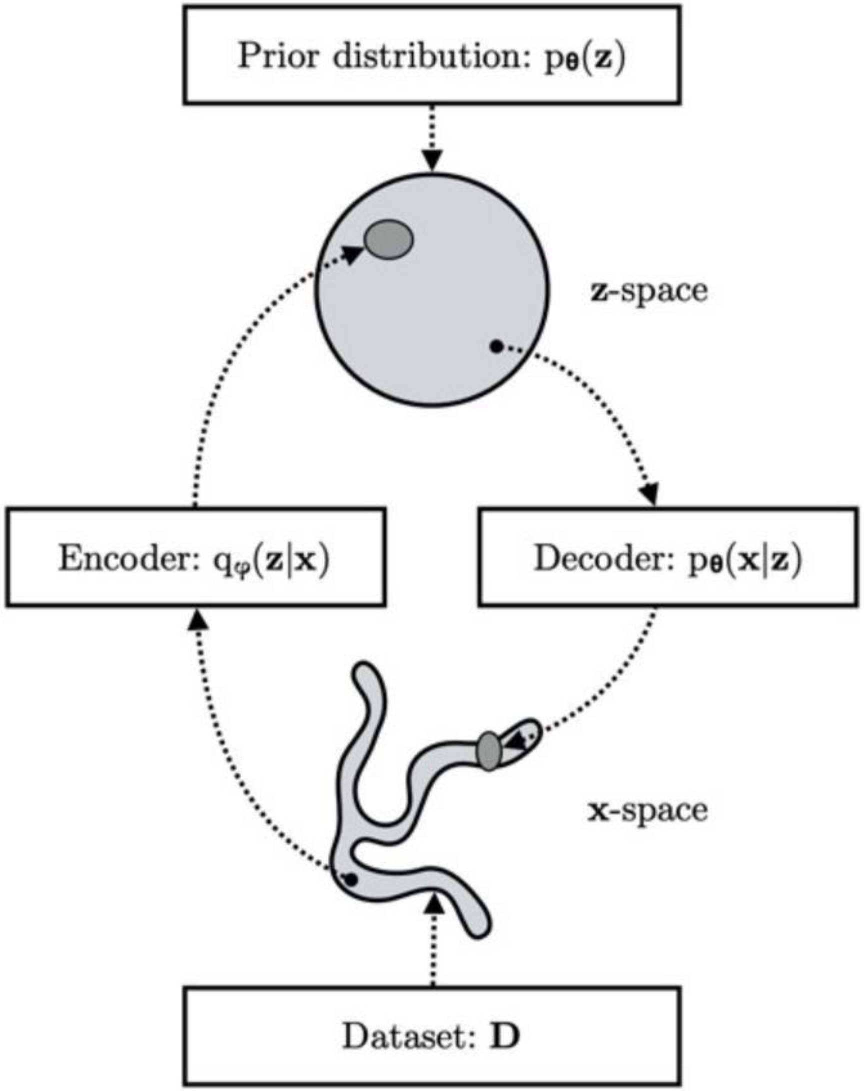


VAE 图结构模型

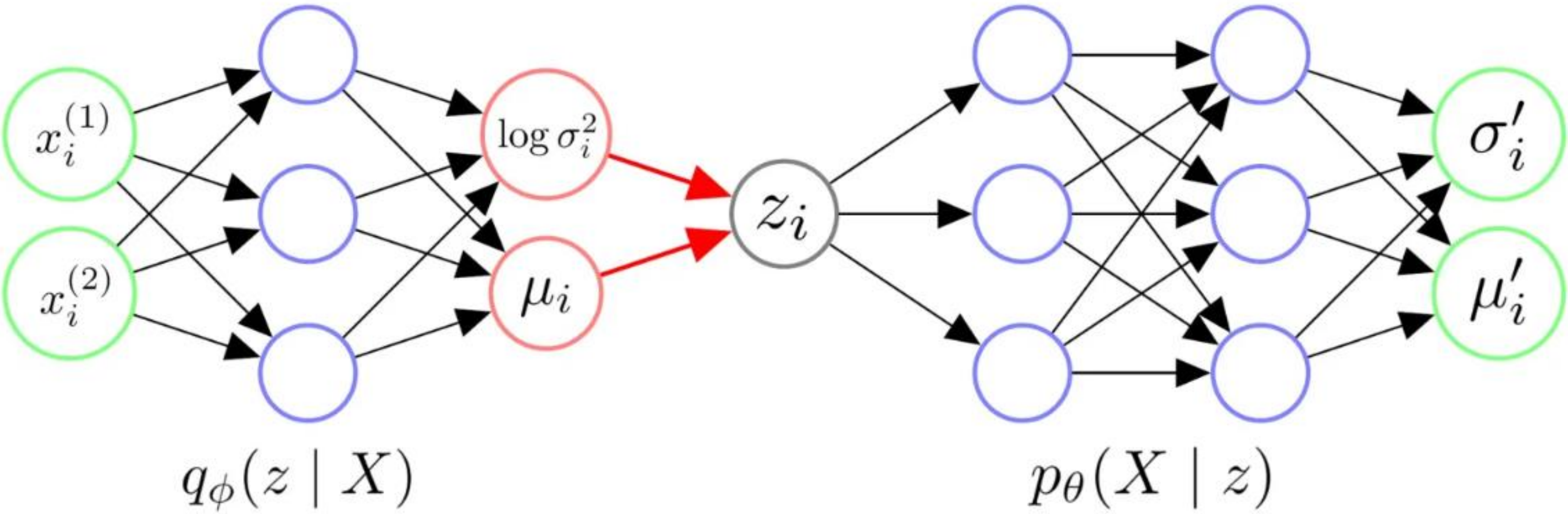


$$P(x) = \int_z P(x|z)P(z)dz$$

VAE的编码和解码过程

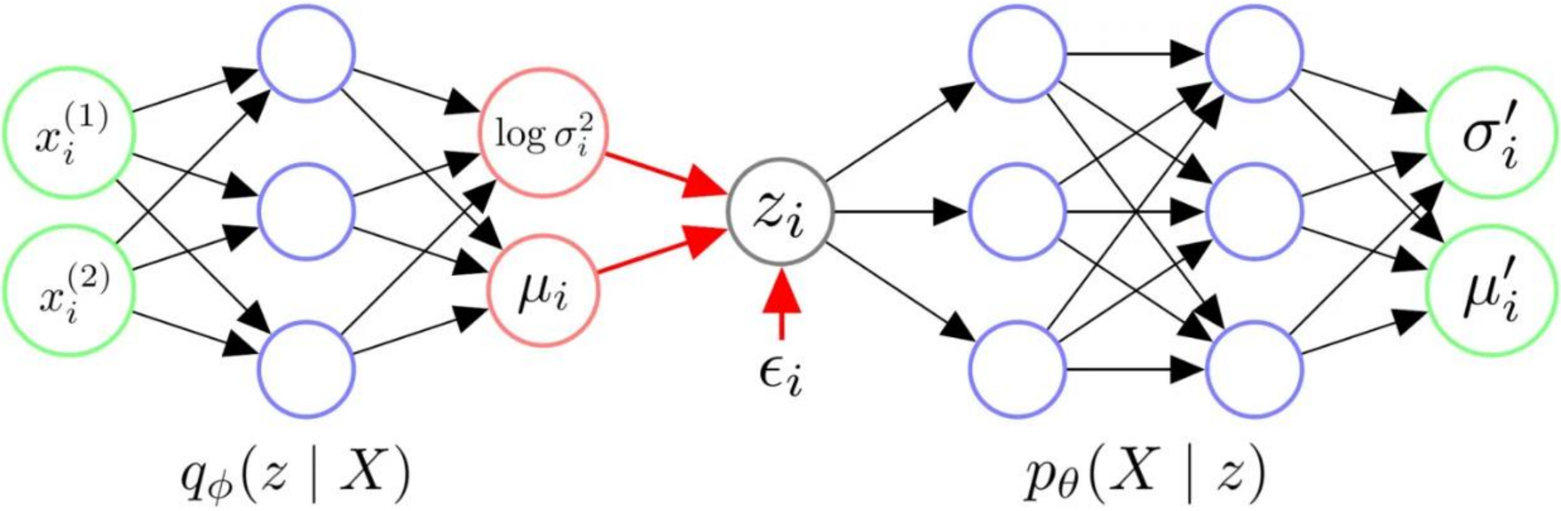


VAE结构



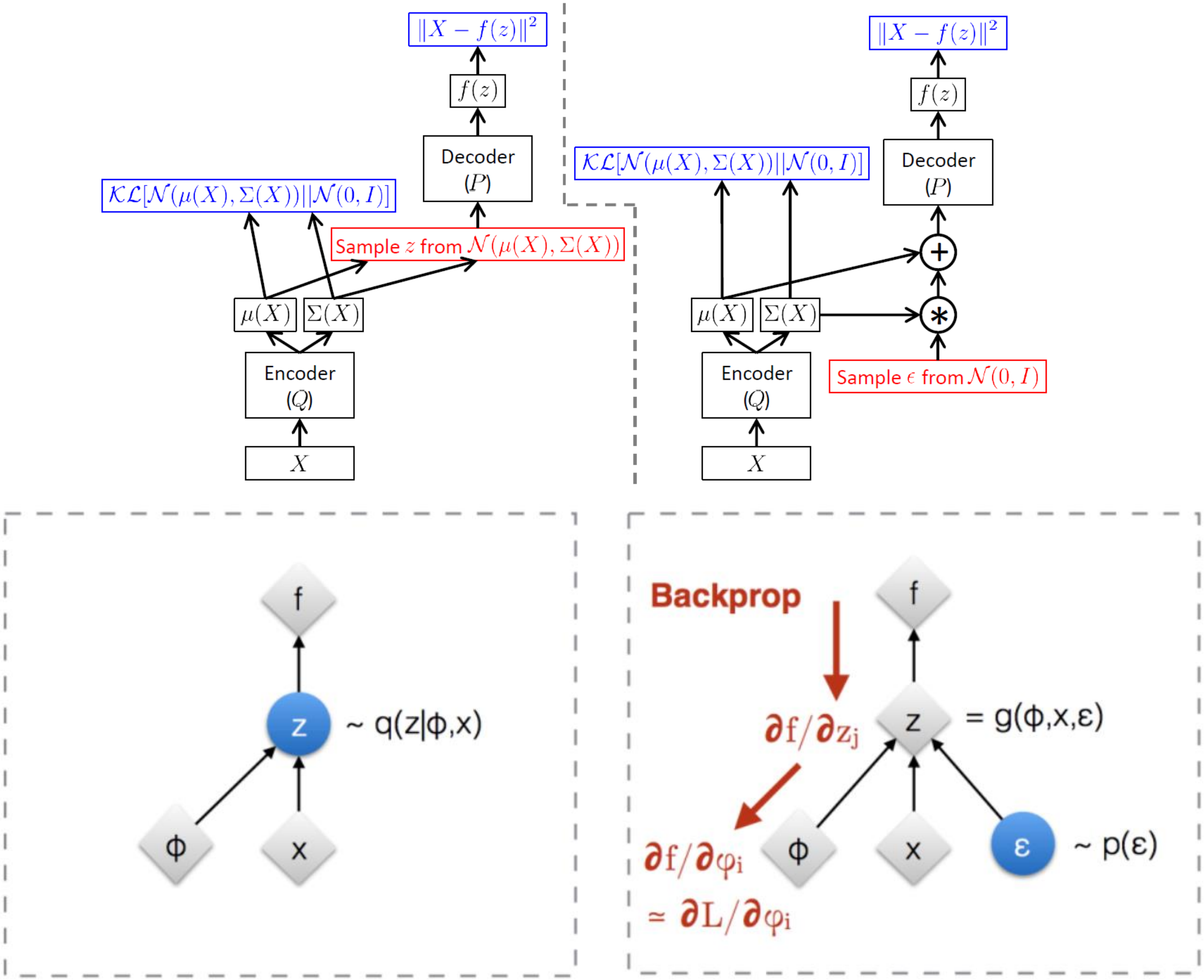
$x_i^{(1)}$
$x_i^{(2)}$
$\log \sigma_i^2$
μ_i
z_i
σ_i'
μ_i'
$q_\phi(z \mid X)$
$p_\theta(X \mid z)$

Reparameterization Trick



$$z_i = \mu_i + \sigma_i \odot \epsilon_i$$

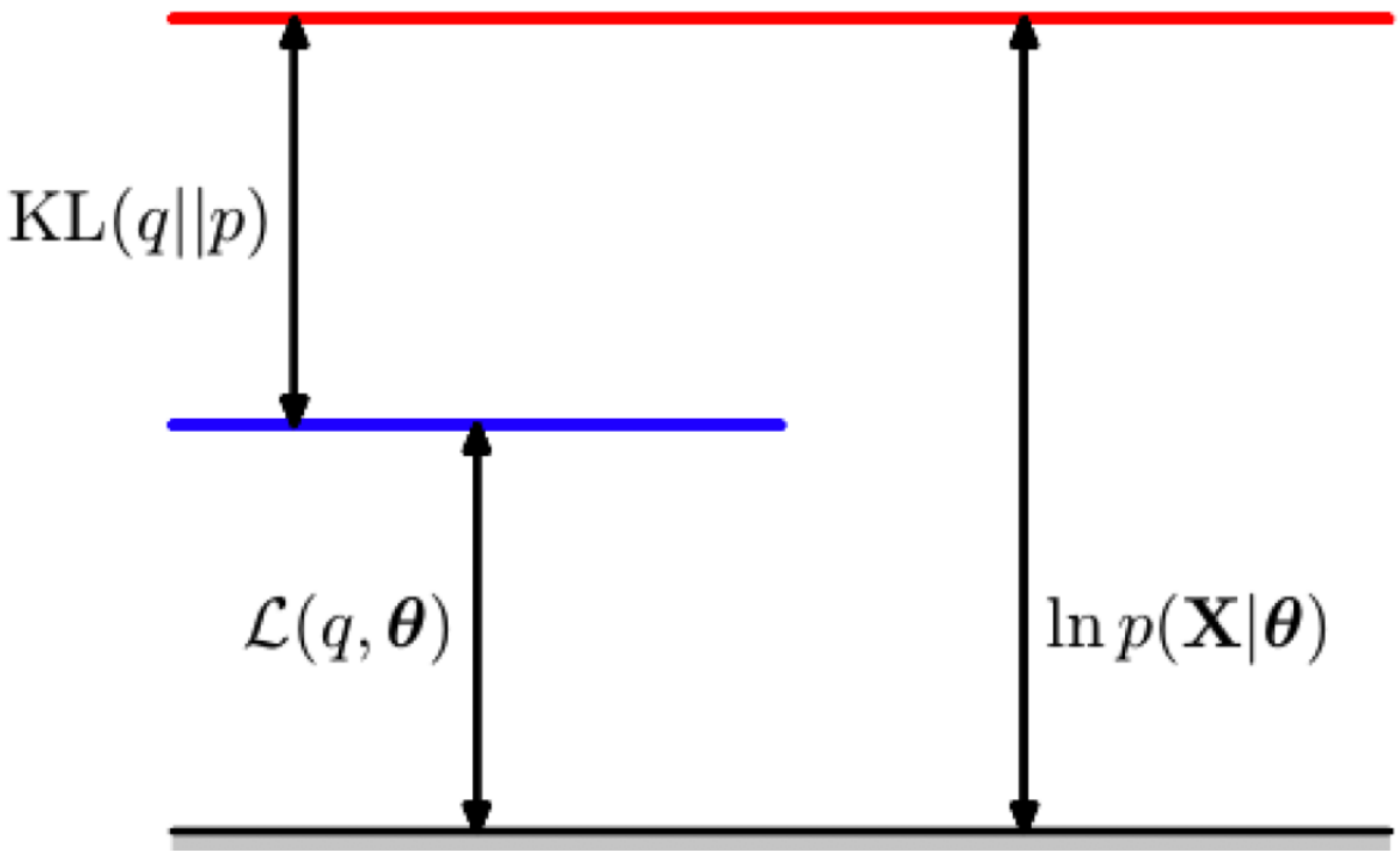
Reparameterization Trick



VAE的ELBO

$$
\begin{aligned}
\log p_\theta(X) &= \int_z q_\phi(z \mid X) \log p_\theta(X) dz \\
&= \int_z q_\phi(z \mid X) \log \frac{p_\theta(X, z)}{p_\theta(z \mid X)} dz \\
&= \int_z q_\phi(z \mid X) \log \left(\frac{p_\theta(X, z)}{q_\phi(z \mid X)} \cdot \frac{q_\phi(z \mid X)}{p_\theta(z \mid X)} \right) dz \\
&= \int_z q_\phi(z \mid X) \log \frac{p_\theta(X, z)}{q_\phi(z \mid X)} dz + \int_z q_\phi(z \mid X) \log \frac{q_\phi(z \mid X)}{p_\theta(z \mid X)} dz \\
&= \ell\left(p_\theta, q_\phi\right) + D_{KL}\left(q_\phi, p_\theta\right) \\
&\geq \ell\left(p_\theta, q_\phi\right)
\end{aligned}
$$

变分推理


$\mathrm{KL}(q||p)$
$\mathcal{L}(q, \boldsymbol{\theta})$
$\ln p(\mathbf{X}|\boldsymbol{\theta})$

优化目标

$$\ell\left(p_\theta, q_\phi\right) = \log p_\theta(X) - D_{KL}\left(q_\phi, p_\theta\right)$$

$$\begin{aligned}
\ell\left(p_\theta, q_\phi\right) &= \int_z q_\phi(z \mid X) \log \frac{p_\theta(X, z)}{q_\phi(z \mid X)} dz \\
&= \int_z q_\phi(z \mid X) \log \frac{p_\theta(X \mid z) p(z)}{q_\phi(z \mid X)} dz \\
&= \int_z q_\phi(z \mid X) \log \frac{p(z)}{q_\phi(z \mid X)} dz + \int_z q_\phi(z \mid X) \log p_\theta(X \mid z) dz \\
&= -D_{KL}\left(q_\phi, p\right) + \mathbb{E}_{q_\phi}\left[\log p_\theta(X \mid z)\right].
\end{aligned}$$

第一项公式

$$
\begin{aligned}
D_{KL}(\mathcal{N}(\mu,\sigma^2)\|\mathcal{N}(0,1)) &= \int_z \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right)\log\frac{\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(z-\mu)^2}{2\sigma^2}\right)}{\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right)}dz \\
&= \int_z \left(\frac{-(z-\mu)^2}{2\sigma^2}+\frac{z^2}{2}-\log\sigma\right)\mathcal{N}(\mu,\sigma^2)\,dz \\
&= -\int_z \frac{(z-\mu)^2}{2\sigma^2}\mathcal{N}(\mu,\sigma^2)\,dz + \int_z \frac{z^2}{2}\mathcal{N}(\mu,\sigma^2)\,dz - \int_z \log\sigma\mathcal{N}(\mu,\sigma^2)\,dz \\
&= -\frac{\mathbb{E}\left[(z-\mu)^2\right]}{2\sigma^2}+\frac{\mathbb{E}\left[z^2\right]}{2}-\log\sigma \\
&= \frac{1}{2}(-1+\sigma^2+\mu^2-\log\sigma^2).
\end{aligned}
$$

$$
D_{KL}(q_\phi(z\mid X),p(z)) = \sum_{j=1}^{d}\frac{1}{2}(-1+\sigma^{(j)^2}+\mu^{(j)^2}-\log\sigma^{(j)^2}).
$$

第二项公式

$$\begin{aligned}\log p_\theta\left(X \mid z_i\right) &= \log \frac{\exp\left(-\frac{1}{2}(X-\mu')^{\mathrm{T}} \Sigma'^{-1}(X-\mu')\right)}{\sqrt{(2\pi)^k |\Sigma'|}} \\ &= -\frac{1}{2}(X-\mu')^{\mathrm{T}} \Sigma'^{-1}(X-\mu') - \log \sqrt{(2\pi)^k |\Sigma'|} \\ &= -\frac{1}{2} \sum_{k=1}^{K} \frac{\left(X^{(k)}-\mu'^{(k)}\right)^2}{\sigma'^{(k)}} - \log \sqrt{(2\pi)^K \prod_{k=1}^{K} \sigma'^{(k)}}.\end{aligned}$$

$$\mathbb{E}_{q_\phi}\left[\log p_\theta(X \mid z)\right] \approx \frac{1}{m} \sum_{i=1}^{m} \log p_\theta\left(X \mid z_i\right),$$

$$z_i \sim q_\phi\left(z \mid x_i\right) = \mathcal{N}\left(z \mid \mu\left(x_i ; \phi\right), \sigma^2\left(x_i ; \phi\right) * I\right)$$

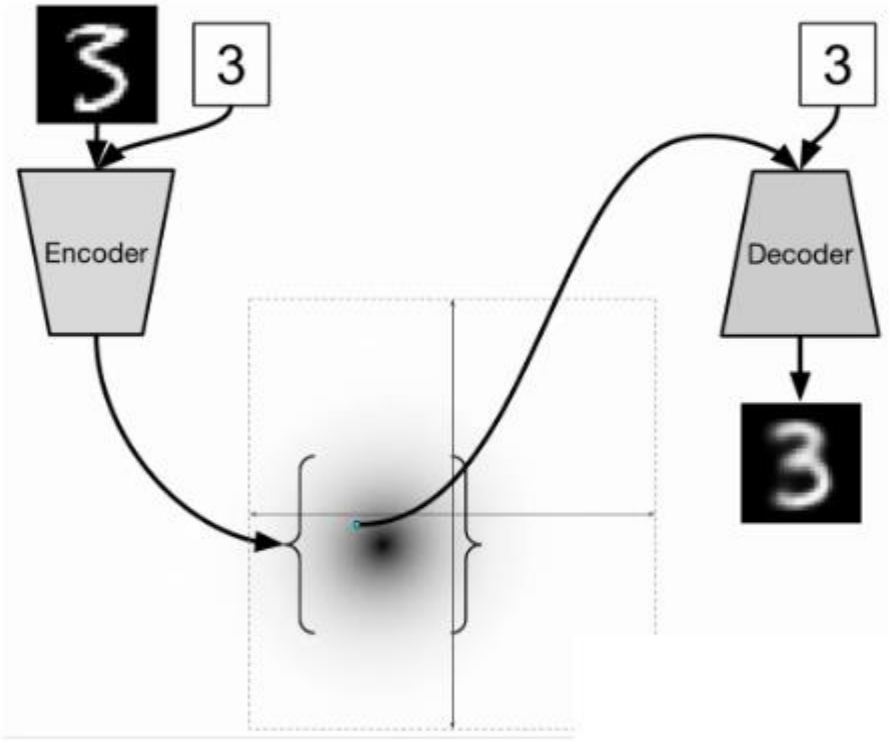
损失函数

$$\begin{aligned}\mathcal{L} &= -\frac{1}{n}\sum_{i=1}^{n}\ell(p_\theta, q_\phi) \\ &= \frac{1}{n}\sum_{i=1}^{n}D_{KL}\left(q_\phi, p\right) - \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{q_\phi}\left[\log p_\theta(x_i \mid z)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}D_{KL}\left(q_\phi, p\right) - \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\log p_\theta\left(x_i \mid z_j\right)\end{aligned}$$

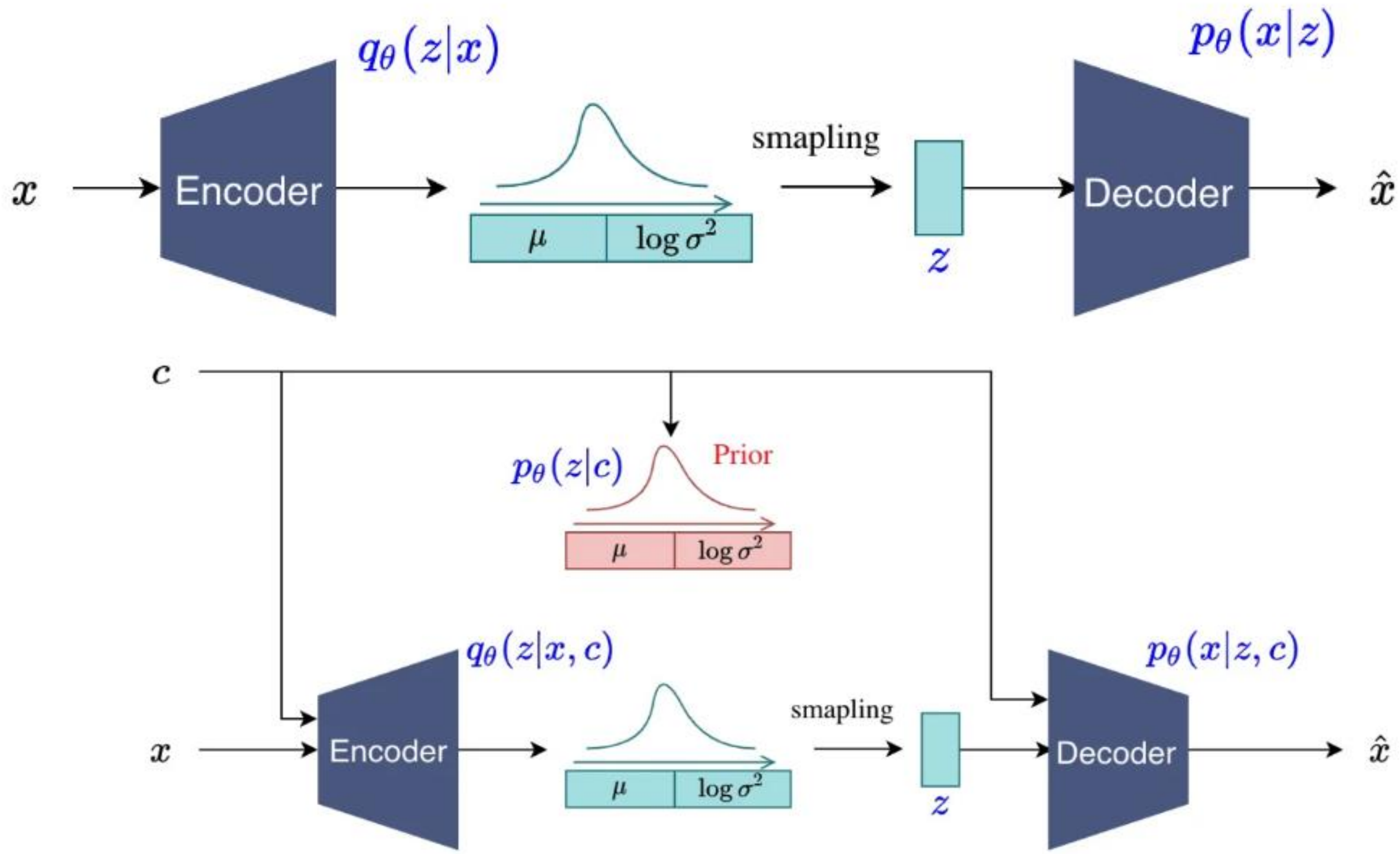
$$\begin{aligned}\mathcal{L} &= \frac{1}{n}\sum_{i=1}^{n}D_{KL}\left(q_\phi, p\right) - \frac{1}{n}\sum_{i=1}^{n}\log p_\theta\left(x_i \mid z_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d}\frac{1}{2}\left(-1 + {\sigma_i^{(j)}}^2 + {\mu_i^{(j)}}^2 - \log {\sigma_i^{(j)}}^2\right) \\ &\quad - \frac{1}{n}\sum_{i=1}^{n}\left(-\frac{1}{2}\sum_{k=1}^{K}\frac{\left(x_i^{(k)} - \mu_i'^{(k)}\right)^2}{\sigma_i'^{(k)}} - \log\sqrt{(2\pi)^K\prod_{k=1}^{K}\sigma_i'^{(k)}}\right)\end{aligned}$$

$$\mathcal{L} = \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\sum_{j=1}^{d}\tfrac{1}{2}\left(-1 + {\sigma_i^{(j)}}^2 + {\mu_i^{(j)}}^2 - \log {\sigma_i^{(j)}}^2\right) + \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\|x_i - \mu_i'\|^2$$

CVAE



模型对比



$q_\theta(z|x)$
$p_\theta(x|z)$
x
Encoder
μ
$\log\sigma^2$
smapling
z
Decoder
$\hat{x}$
c
$p_\theta(z|c)$
Prior
μ
$\log\sigma^2$
$q_\theta(z|x,c)$
$p_\theta(x|z,c)$
x
Encoder
μ
$\log\sigma^2$
smapling
z
Decoder
$\hat{x}$

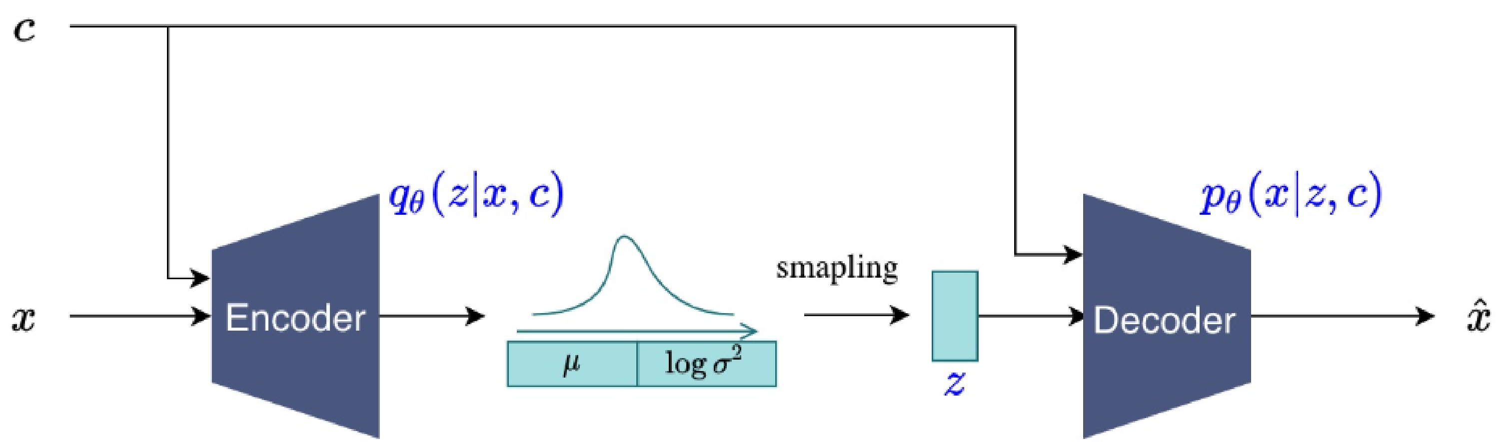
优化目标对比

$$
\begin{aligned}
\log p(x) &= \mathbb{E}_{q(z|x)}[\log p(x)] \\
&= \mathbb{E}_{q(z|x)}[\log \frac{p(x,z)}{p(z|x)}] = \mathbb{E}_{q(z|x)}[\log \frac{q(z|x)p(x,z)}{p(z|x)q(z|x)}] \\
&= \mathbb{E}_{q(z|x)}[\log p(x,z) - \log q(z|x)] + \underbrace{D_{KL}(q(z|x)\|p(z|x))}_{\geq 0} \\
&\geq \mathbb{E}_{q(z|x)}[\log p(x,z) - \log q(z|x)] \\
&:= ELBO \\
&= \mathbb{E}_{q(z|x)}[\log p(z) + \log p(x|z) - \log q(z|x)] \\
&= \underbrace{\mathbb{E}_{q(z|x)}[\log p(x|z)]}_{\text{Reconstruct term } L_{Rec}} - \underbrace{D_{KL}(q(z|x)\|p(z))}_{\text{KL term } L_{KL}}
\end{aligned}
$$

$$
\begin{aligned}
\log p(x|c) &= \mathbb{E}_{q(z|x,c)}[\log p(x|c)] \\
&= \mathbb{E}_{q(z|x,c)}[\log \frac{p(x,z|c)}{p(z|x,c)}] \\
&= \mathbb{E}_{q(z|x,c)}[\log \frac{p(x,z|c)}{q(z|x,c)} \frac{q(z|x,c)}{p(z|x,c)}] \\
&= \mathbb{E}_{q(z|x,c)}[\log p(x,z|c) - \log q(z|x,c)] + \underbrace{D_{KL}(q(z|x,c)\|p(z|x,c))}_{\geq 0} \\
&\geq \mathbb{E}_{q(z|x,c)}[\log p(x,z|c) - \log q(z|x,c)] \\
&:= ELBO \\
&= \mathbb{E}_{q(z|x,c)}[\log p(z|c) + \log p(x|z,c) - \log q(z|x,c)] \\
&= \underbrace{\mathbb{E}_{q(z|x,c)}[\log p(x|z,c)]}_{\text{Reconstruct term } L_{Rec}} - \underbrace{D_{KL}(q(z|x,c)\|p(z|c))}_{\text{KL term } L_{KL}}
\end{aligned}
$$

简化版CVAE

$$\log p(x|c) = \underbrace{\mathbb{E}_{q(z|x,c)}[\log p(x|z,c)]}_{\text{Reconstruct term } L_{Rec}} - \underbrace{D_{KL}(q(z|x,c)\|p(z))}_{\text{KL term } L_{KL}}$$



VAE的不足

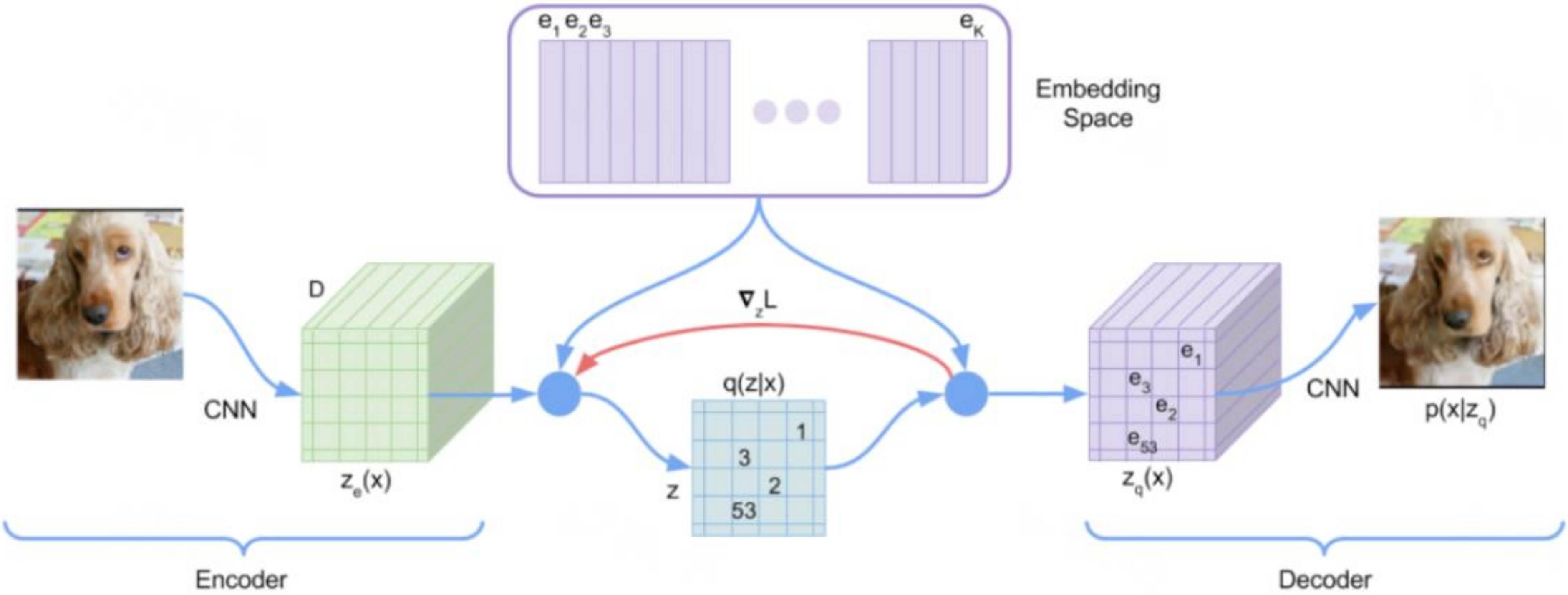
- VAE与AE不同之处在与，VAE不再去学习一个连续的表征，而是直接学习一个分布，然后通过这个分布采样得到中间表征去重建原图。

- VAE的优化目标为

$$
\begin{aligned}
\mathcal{L} &= \frac{1}{n}\sum_{i=1}^{n} D_{KL}\left(q_\phi, p\right) - \frac{1}{n}\sum_{i=1}^{n} \log p_\theta\left(x_i \mid z_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{d}\frac{1}{2}\left(-1+\sigma_i^{(j)^2}+\mu_i^{(j)^2}-\log\sigma_i^{(j)^2}\right) \\
&\quad -\frac{1}{n}\sum_{i=1}^{n}\left(-\frac{1}{2}\sum_{k=1}^{K}\frac{\left(x_i^{(k)}-\mu_i^{\prime(k)}\right)^2}{\sigma_i^{\prime(k)}}-\log\sqrt{(2\pi)^K\prod_{k=1}^{K}\sigma_i^{\prime(k)}}\right)
\end{aligned}
$$

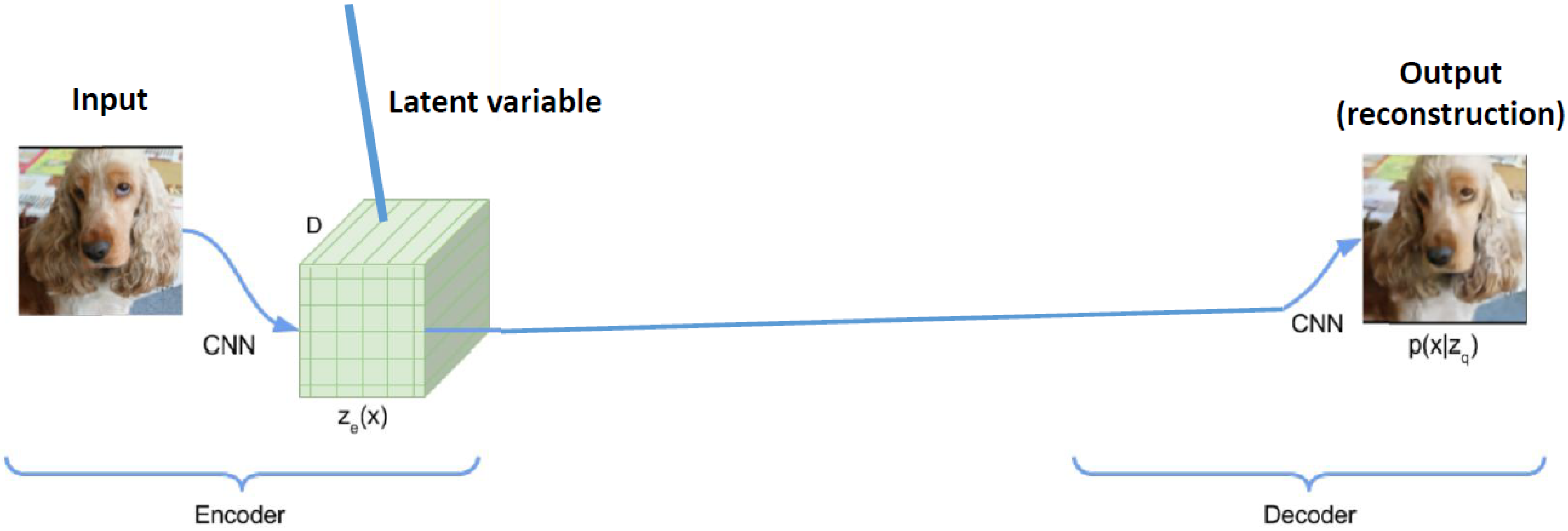
- VAE使用了固定的正态分布先验，以及连续的中间表征，导致图片生成的多样性弱和可控性差。

VQ-VAE



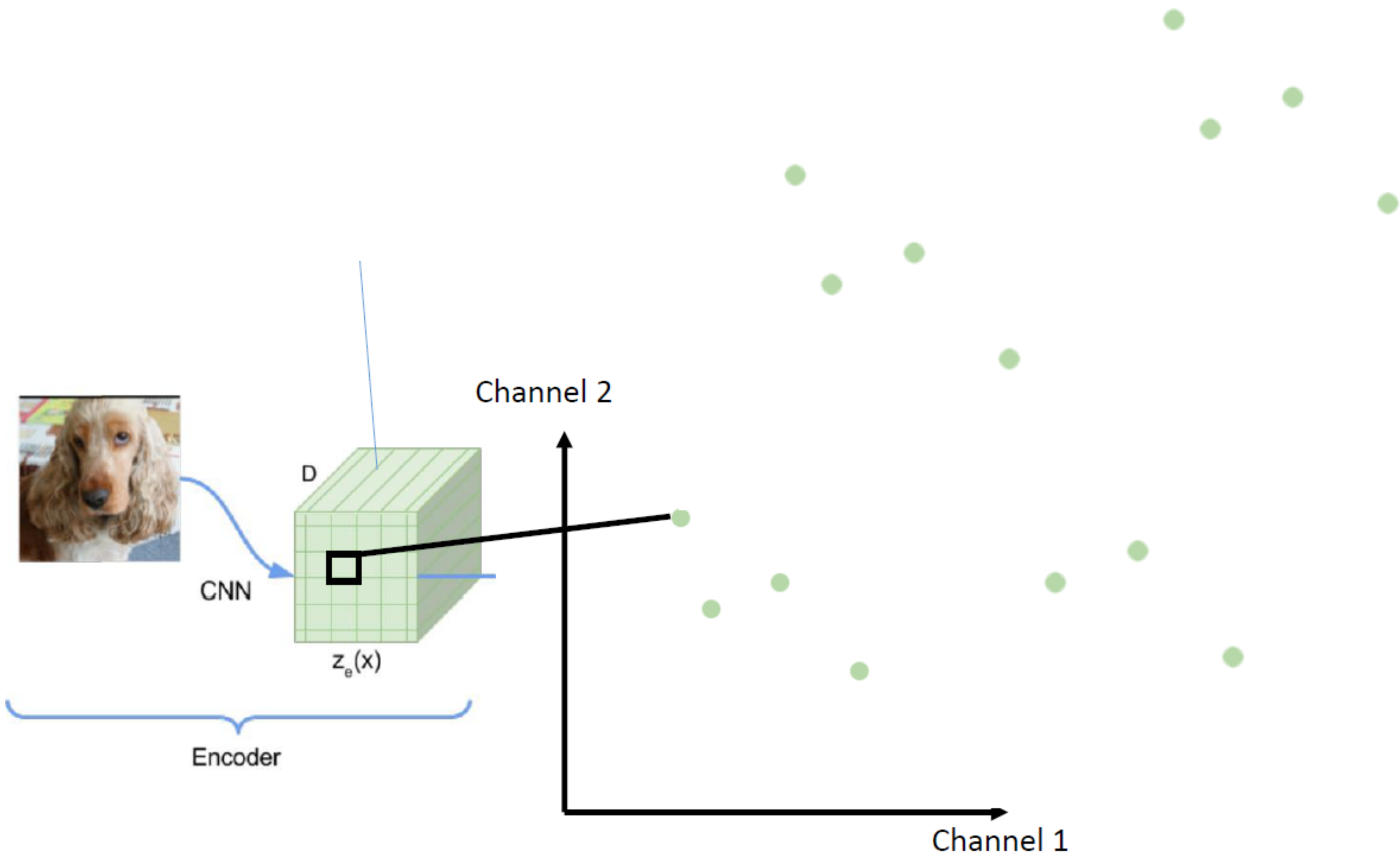
$e_1 e_2 e_3$
e_K
Embedding Space
D
$\nabla_z L$
CNN
$z_e(x)$
q(z|x)
1
3
2
53
z
e_1
e_3
e_2
e_{53}
$z_q(x)$
CNN
$p(x|z_q)$
Encoder
Decoder

自动编码

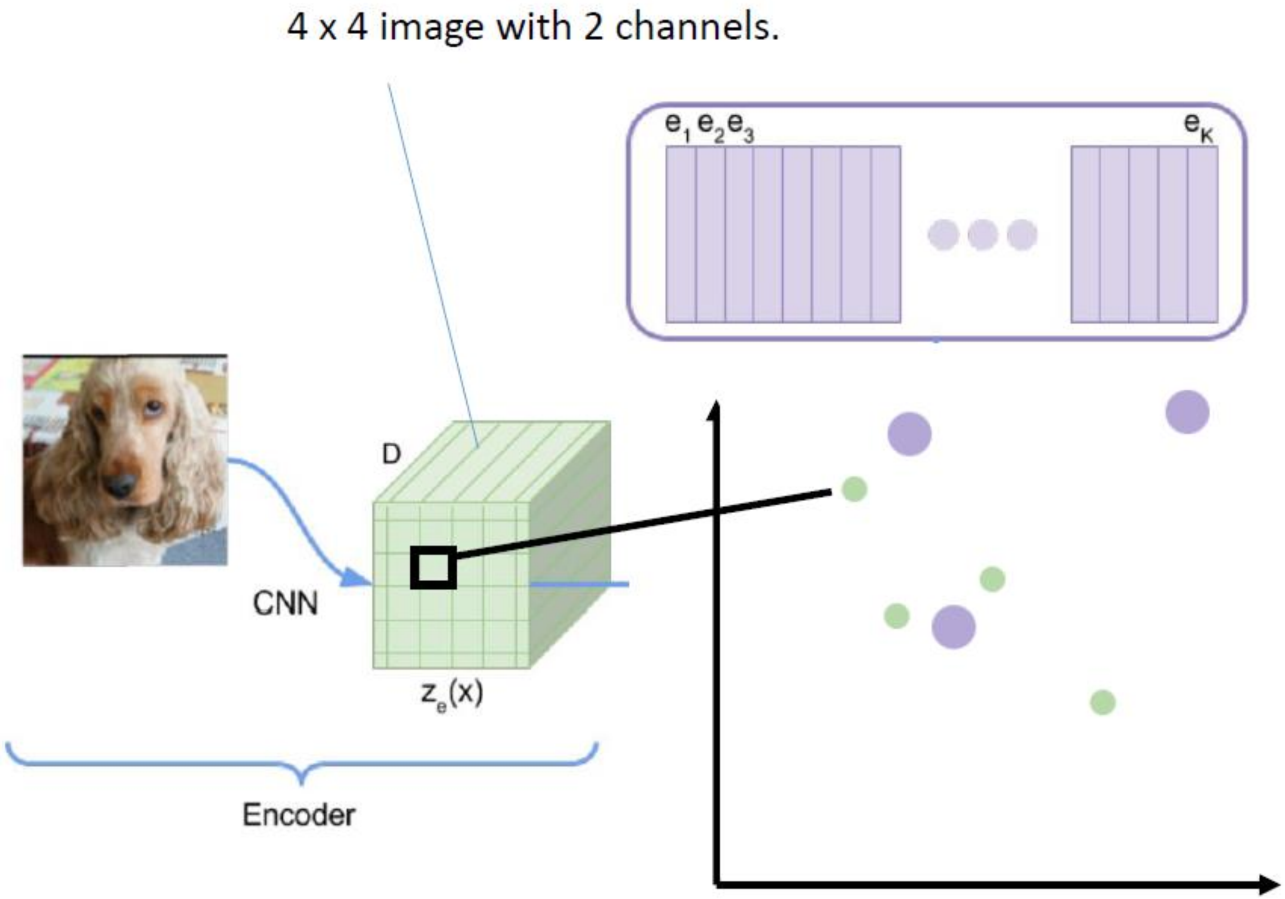


Input
Latent variable
Output
(reconstruction)
D
CNN
$z_e(x)$
CNN
$p(x|z_q)$
Encoder
Decoder

编码器输出表示



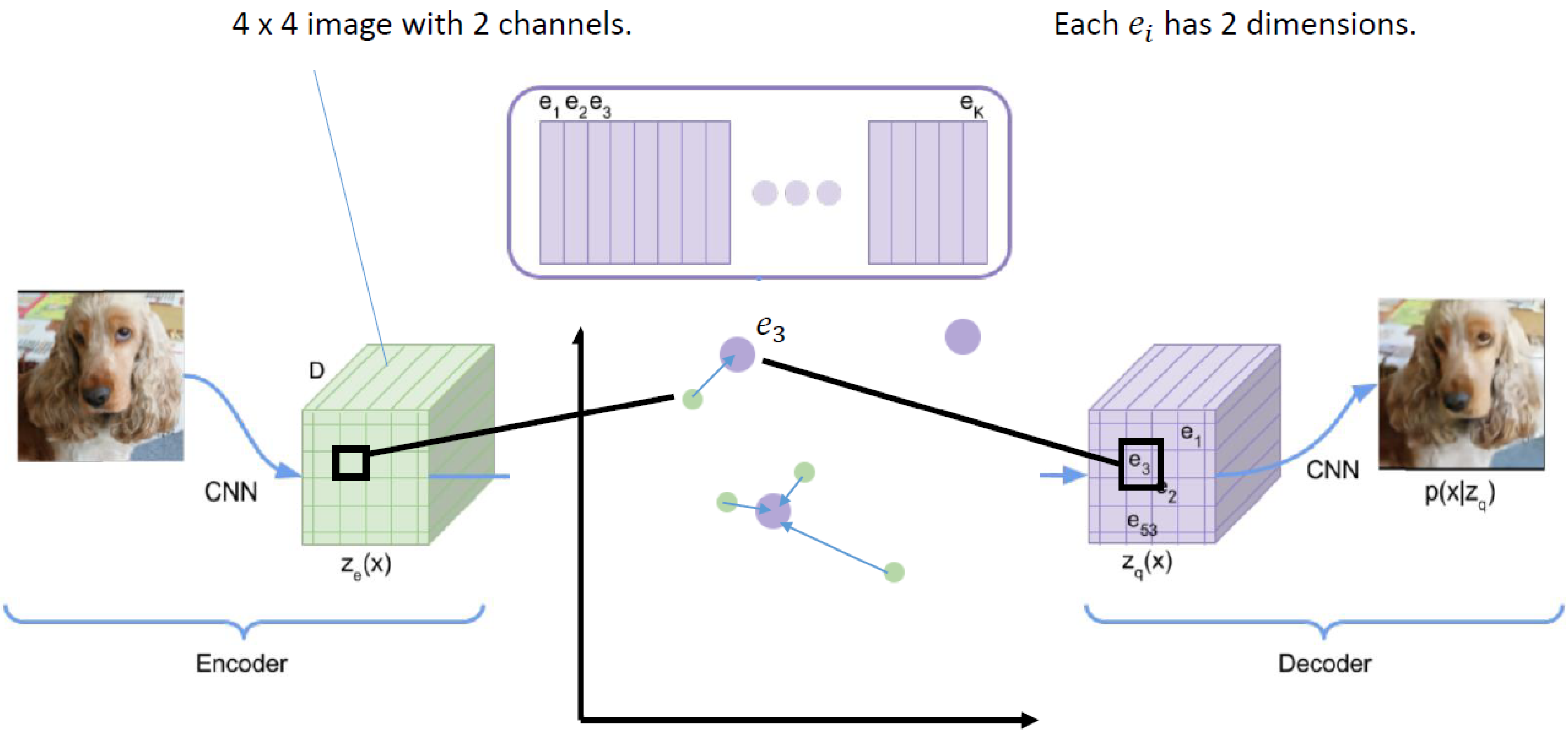
降维离散化



Make dictionary of vectors $e_1, \ldots, e_K$

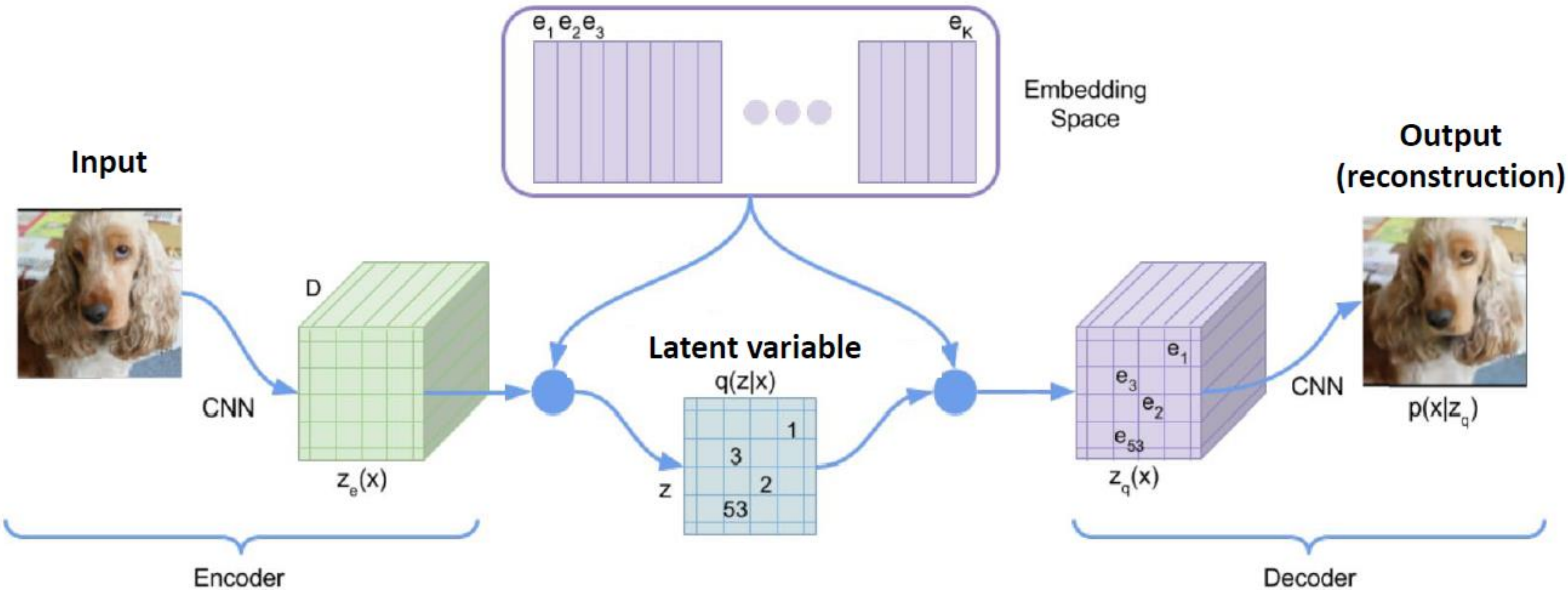
Each e_i has 2 dimensions.

最近邻重构



4 x 4 image with 2 channels.
Each e_i has 2 dimensions.
e1 e2 e3
eK
e3
D
CNN
z_e(x)
e1
e3
e2
e53
z_q(x)
CNN
p(x|z_q)
Encoder
Decoder

VQ-VAE模型



e_1 e_2 e_3
e_K
Embedding Space
Input
Output (reconstruction)
D
Latent variable
q(z|x)
1
3
2
53
z
e_1
e_3
e_2
e_{53}
CNN
CNN
$z_e(x)$
$z_q(x)$
$p(x|z_q)$
Encoder
Decoder

最近邻重构

- 重构流程是：

$$z = encoder(x)$$

$$E = [e_1, e_2, \ldots, e_K]$$

$$z \to e_k, \quad k = \arg\min_j \|z - e_j\|_2$$

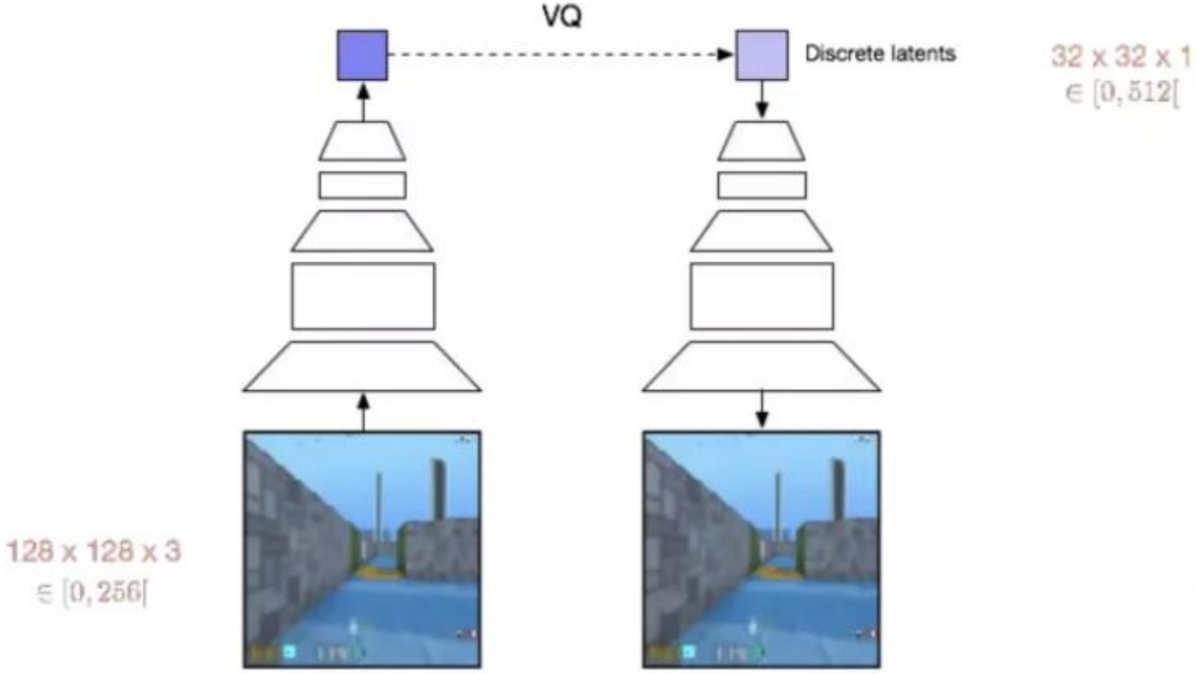
$$x \xrightarrow{encoder} z \xrightarrow{\text{最邻近}} z_q \xrightarrow{decoder} \hat{x}$$

$$z = \begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1m} \\ z_{21} & z_{22} & \cdots & z_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ z_{m1} & z_{m2} & \cdots & z_{mm} \end{pmatrix}$$

向量量化（VQ）

$$z_q(x) = \text{Quantize}(z_e(x)) = e_k \text{ where } k = \arg\min_i \|z_e(x) - \mathbf{e}_i\|_2$$

$$x \rightarrow z_e(x) = \text{Encoder}(x) \rightarrow z_q(x) = \text{Quantize}(z_q(x)) \rightarrow x' = \text{Decoder}(z_q(x))$$



后验分布 q(z|x)

- 后验分布 q(z|x) 是一个多类分布（categorical distribution），其概率分布为one-hot类型：

$$q(z=k|x) = \begin{cases} 1 & \text{for } k = \arg\min_i \|z_e(x) - \mathbf{e}_i\|_2 \\ 0 & \text{otherwise} \end{cases}$$

- 基确定分布q(z|x), 后验分布q(z|x)和先验分布p(z)的KL散度为：

$$\begin{aligned} \mathrm{KL}(q(z|x)||p(z)) &= \sum q(z|x) \log \frac{q(z|x)}{p(z)} \\ &= 1 \cdot \log \frac{1}{1/K} + (K-1) \cdot 0 \cdot \log \frac{0}{1/K} \\ &= \log K \end{aligned}$$

- 给定KL散度为一个常量，VQ-VAE的训练损失项为重建误差 log p(x|z)。

VQ-VAE目标函数

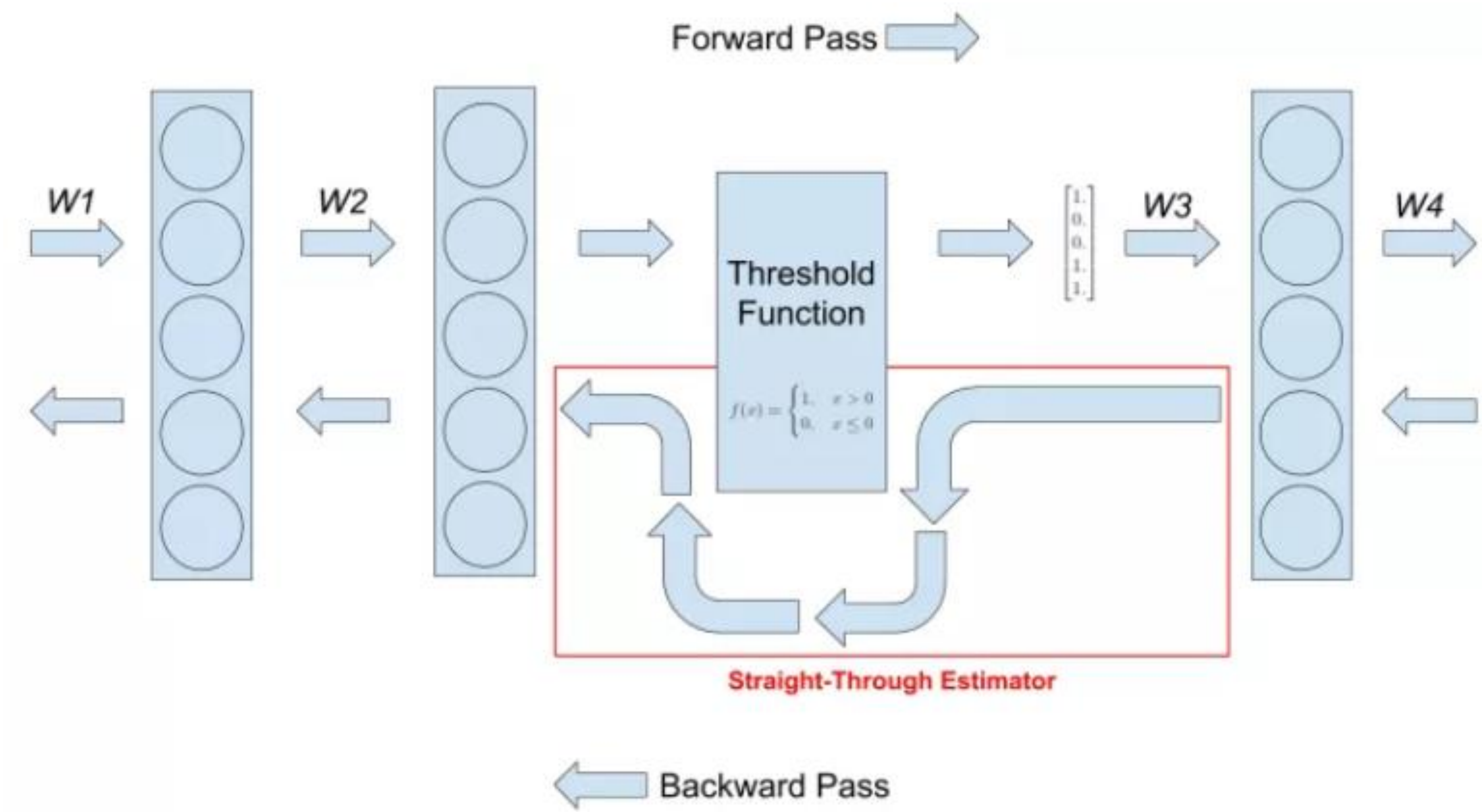
- VQ-VAE的目标函数包含三个部分的训练损失：reconstruction loss, VQ loss, commitment loss:

$$L = \underbrace{\log p(x|z_q(x))}_{\text{reconstruction loss}} + \underbrace{\|\text{sg}[z_e(x)] - e_k\|_2^2}_{\text{VQ loss}} + \underbrace{\beta\|z_e(x) - \text{sg}[e_k]\|_2^2}_{\text{commitment loss}}$$

- 其中，reconstruction loss作用在encoder和decoder上，VQ loss用来更新embedding空间（EMA方式），commitment loss用来约束encoder。系数beta默认设置为0.25。

Straight-through Estimator

- 由于argmin操作不可导，重建误差的梯度无法传导到encoder, 采用straight-through estimator来采用上游得到的梯度。



自行设计梯度

- 基于Straight-Through的思想，前项传播的时候可以用想要的变量，而反向传播的时候，用所涉及的梯度。其目标函数为

$$\|x - decoder(z + sg[z_q - z])\|_2^2$$

- 其中，前向传播计算为：

$$decoder(z + z_q - z) = decoder(z_q)$$

VQ损失项学习

- 采用EMA (Exponential moving averages)来更新量化向量：

$$\|\text{sg}[z_e(x)] - e\|_2^2$$

$$\sum_j^{n_i} \|z_{i,j} - e_i\|_2^2 \qquad \{z_{i,1}, z_{i,2}, \ldots, z_{i,n_i}\}$$

$$e_i = \frac{1}{n_i} \sum_j^{n_i} z_{i,j}$$

$$N_i^{(t)} := N_i^{(t-1)} * \gamma + n_i^{(t)}(1-\gamma)$$

$$m_i^{(t)} := m_i^{(t-1)} * \gamma + \sum_j z_{i,j}^{(t)}(1-\gamma)$$

$$e_i^{(t)} := \frac{m_i^{(t)}}{N_i^{(t)}},$$

Commitment损失项学习

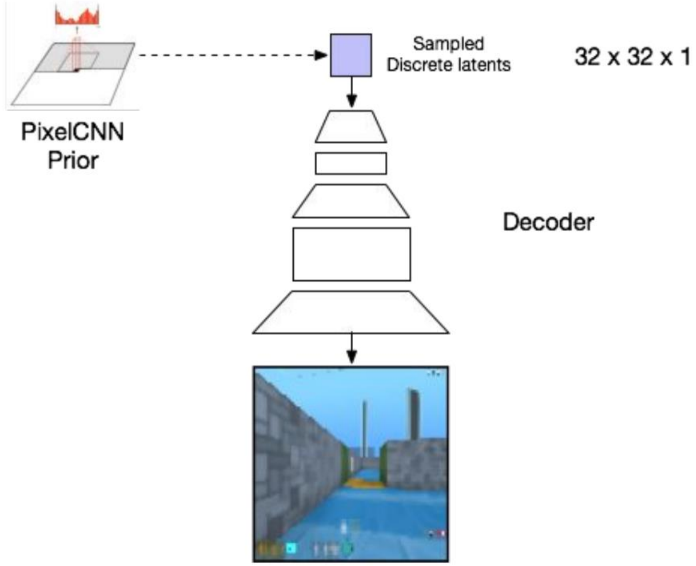
- Commitment训练损失项主要约束编码器（encoder）的输出和量化向量（embedding）空间保持一致，避免encoder的输出变动较大。

- Commitment损失计算encoder的输出和对应的量化得到的embedding的向量l2误差，仅影响encoder。

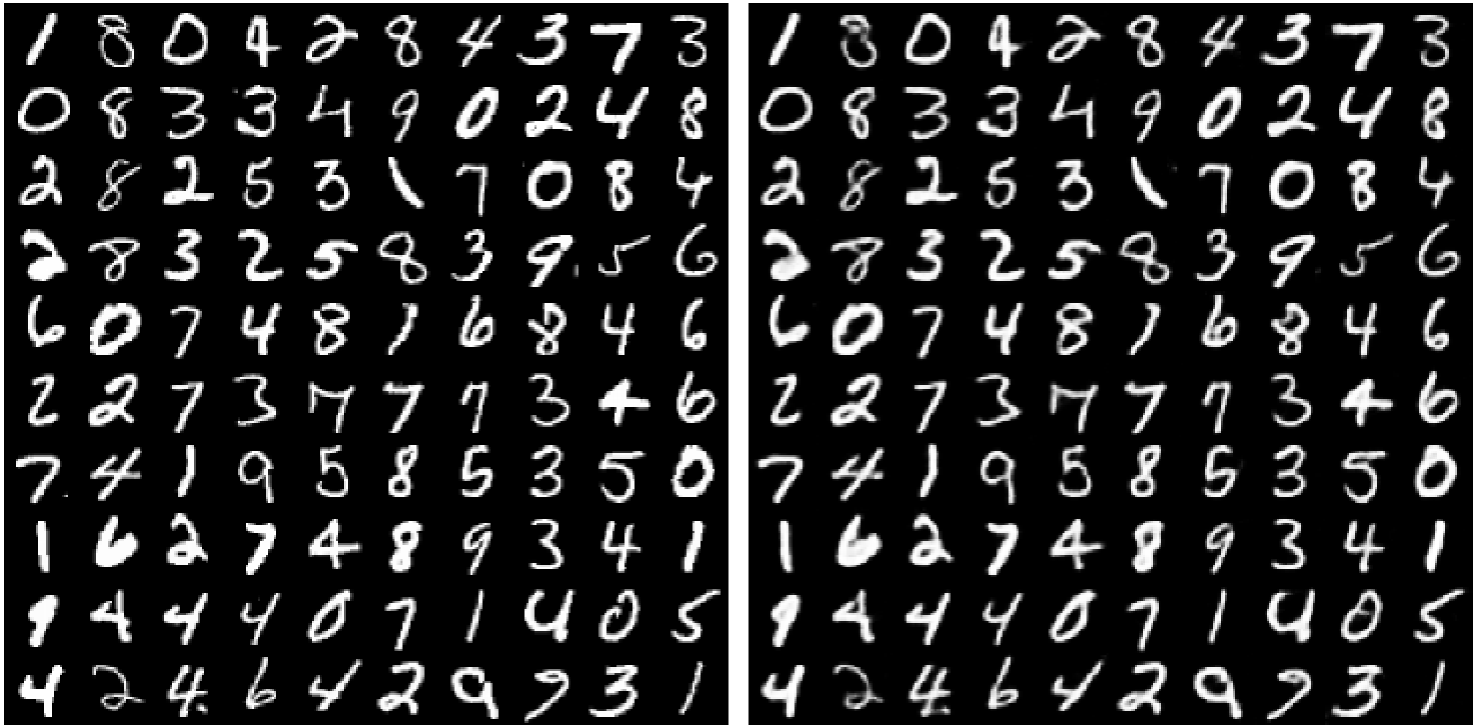
$$\|z_e(x) - \mathrm{sg}[e_k]\|_2^2$$

拟合编码分布

- 利用自回归模型PixelCNN，对编码矩阵进行拟合，从而构建先验分布。
- 通过PixelCNN得到编码分布后，随机生成一个新的编码矩阵，然后通过编码表映射为量化矩阵，最后经过decoder得到一张图片。



MINIST实验结果（训练集）



MINIST实验结果（测试集）