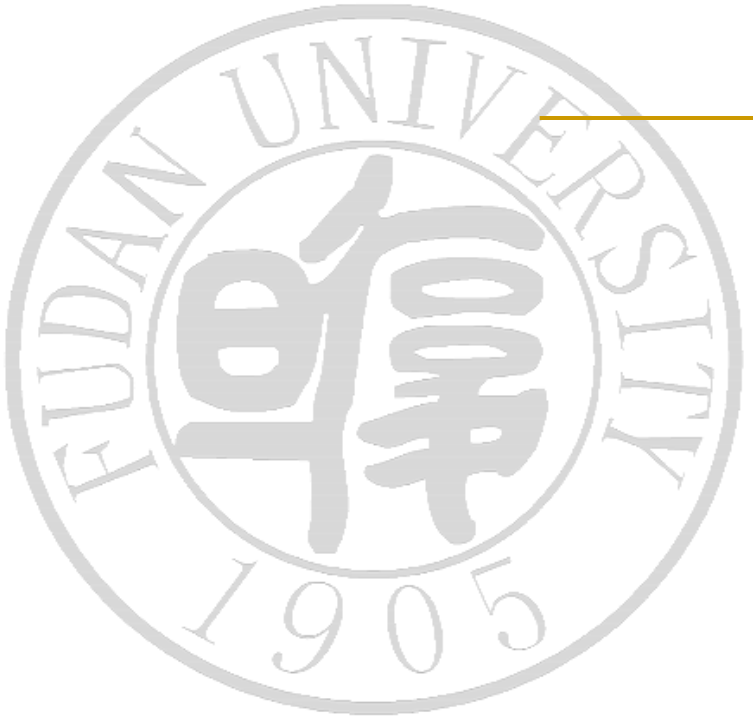
Generative Models: Fundamentals and Applications

Lecture 8: Variational Autoencoders



Shuigeng Zhou, Yuxi Mi
College of CSAI

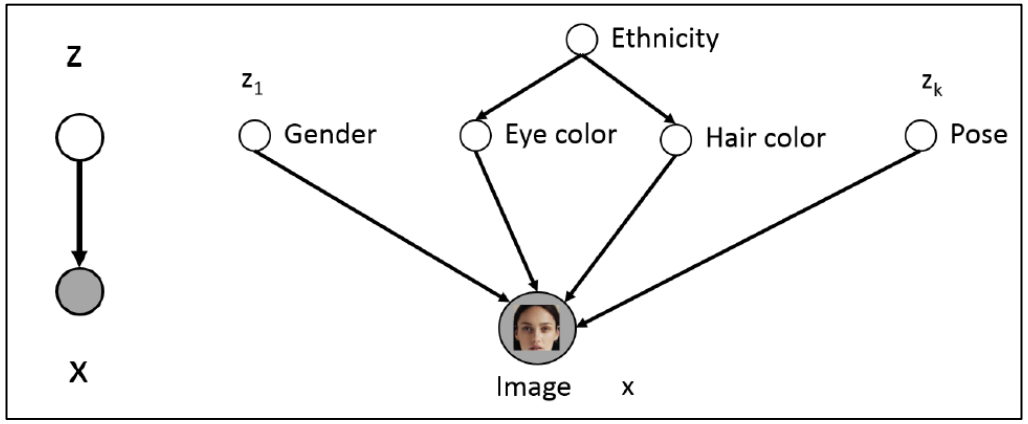
November 24, 2025

目录

- 隐变量模型
- 自动编码器(Auto-Encoder)
- 变分自编码器(Variational Auto-Encoder)
- 最新进展

隐变量模型

- 实例：人脸
  - 性别、眼睛颜色、头发颜色、人脸姿势等特征没有通过 x 显式表示
  - 用隐变量 z 表示这些特征

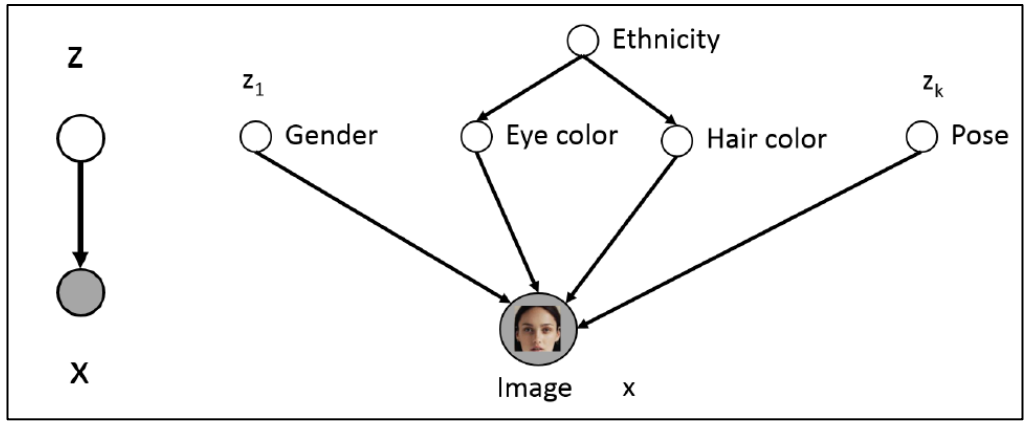


隐变量模型

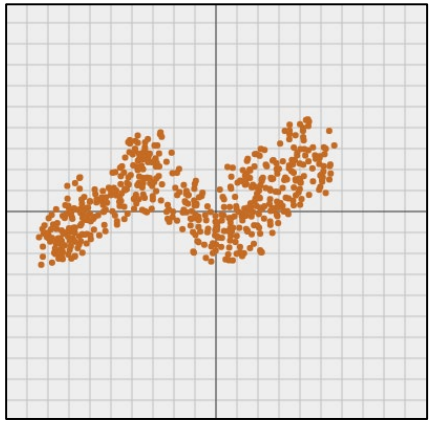
- 条件概率 $p(x|z)$
  - 如果z选的好，那么$p(x|z)$比$p(x)$简单

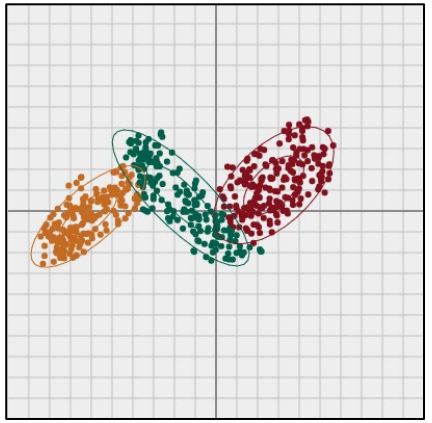
$$p(x) = \int_z p(z)p(x|z)\,dz$$

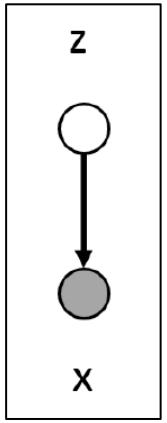
- 条件概率 $p(z|x)$
  - 给定数据，确定其隐含特征



实例：高斯混合模型







- $z \sim \mathrm{Cat}(1, \ldots, K)$
- 生成，从对应分布采样 $p(x|z = k) \sim \mathcal{N}(\mu_k, \Sigma_k)$
- 推断，后验分布 $p(z|x)$
  - 识别数据x是由哪个高斯模型生成的

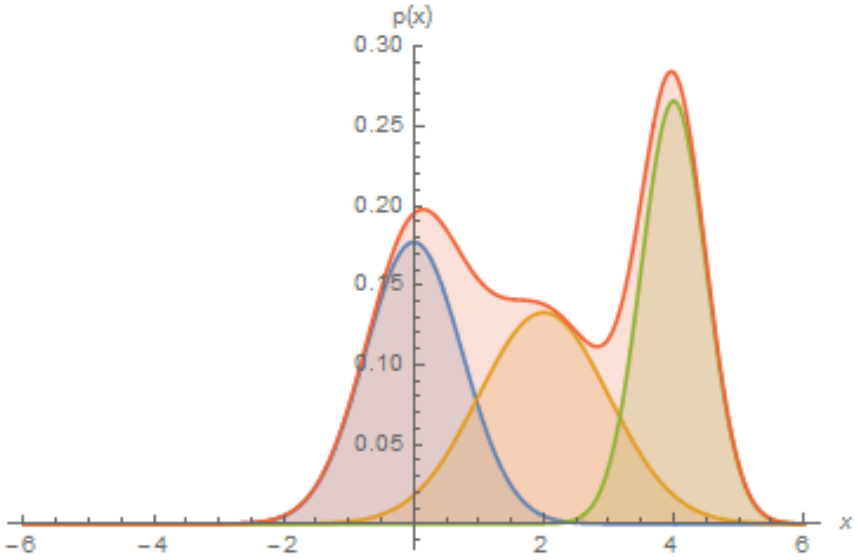
实例：高斯混合模型

- 生成过程
  - 选择 $z = k$
  - 从第 k 个高斯分布中采样，生成数据 x
  - 模型对 x 的整体概率密度

$$p(x) = \sum_z p(x, z) = \sum_z p(z)p(x|z)$$

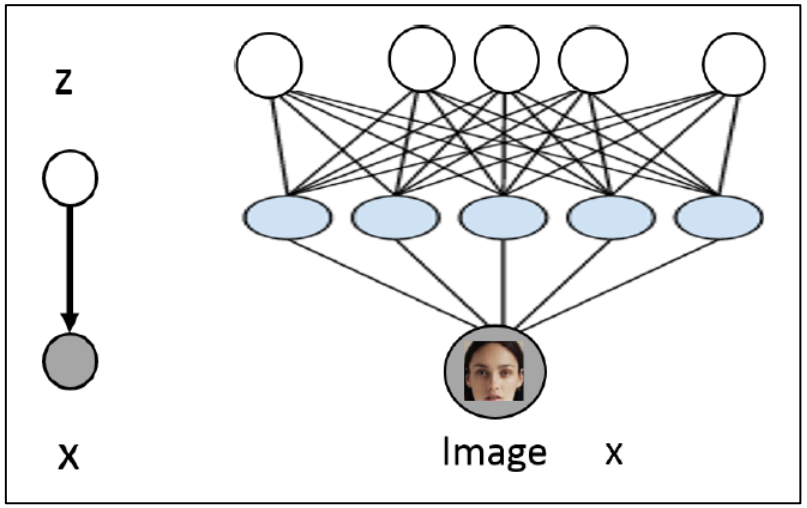
$$= \sum_{k=1}^{K} p(z = k)p(x|z = k)$$

$$= \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$$



深度隐变量模型

- 隐变量连续
  - $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{1})$
- $p(x|z) \sim \mathcal{N}\left(\mu_\theta(z), \Sigma_\theta(z)\right)$
  - $\mu_\theta(z), \Sigma_\theta(z)$为神经网络
  - 模型参数$\theta = (A, B, c, d)$



$$\mu_\theta(\mathbf{z}) = \sigma(A\mathbf{z} + c) = (\sigma(a_1\mathbf{z} + c_1), \sigma(a_2\mathbf{z} + c_2)) = (\mu_1(\mathbf{z}), \mu_2(\mathbf{z}))$$

$$\Sigma_\theta(\mathbf{z}) = diag(\exp(\sigma(B\mathbf{z} + d))) = \begin{pmatrix} \exp(\sigma(b_1\mathbf{z}+d_1)) & 0 \\ 0 & \exp(\sigma(b_2\mathbf{z}+d_2)) \end{pmatrix}$$

深度隐变量模型

- 尽管条件概率$p(x|z)$非常简单，分布$p(x)$很复杂

$$p(x) = \int_z p(z)p(x|z)\,dz$$

- 极大似然估计

$$\log \prod_{\mathbf{x}\in\mathcal{D}} p(\mathbf{x};\theta) = \sum_{\mathbf{x}\in\mathcal{D}} \log p(\mathbf{x};\theta) = \sum_{\mathbf{x}\in\mathcal{D}} \log \sum_{\mathbf{z}} p(\mathbf{x},\mathbf{z};\theta)$$

  - EM算法、变分推断

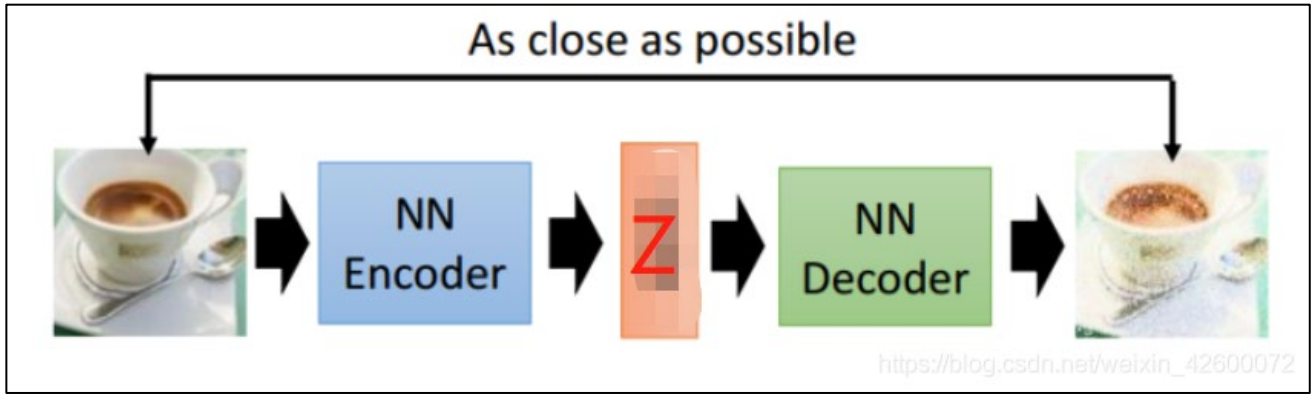
目录

- 隐变量模型
- 自动编码器(Auto-Encoder, AE)
- 变分自编码器(Variational Auto-Encoder, VAE)
- 最新进展

自动编码器

- 传统的自动编码器是一种数据的压缩算法，其算法包括编码阶段和解码阶段，且拥有对称的结构。

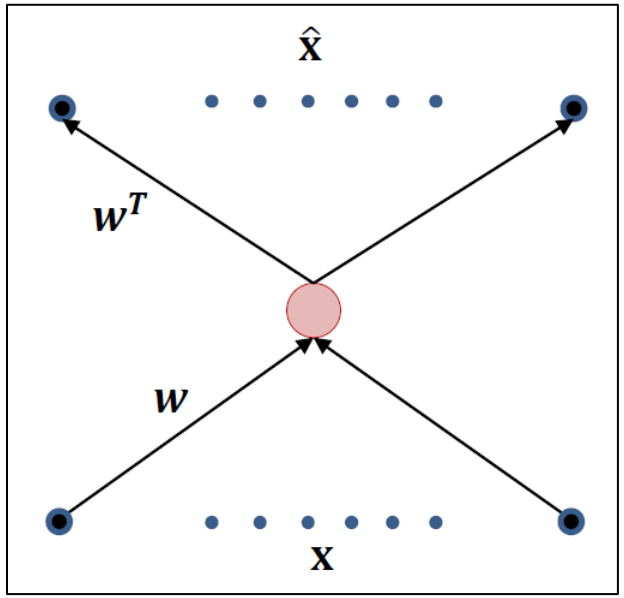
- 应用
  - 数据去噪
  - 数据降维



最简单的自动编码器

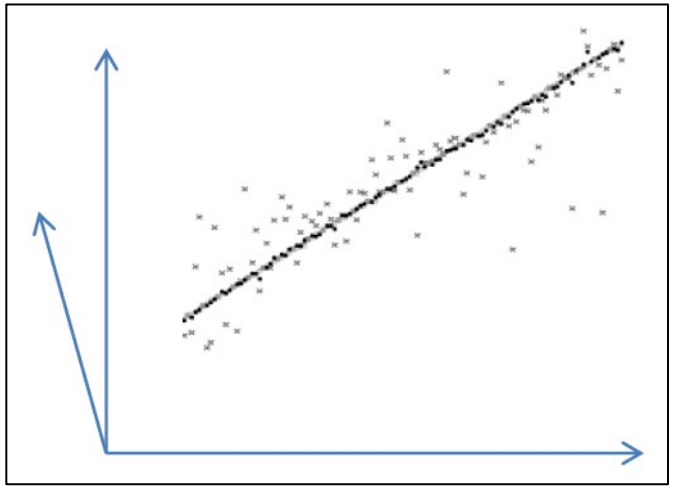
- Toy Model：只有1个隐藏单元
  - 使用同一个参数矩阵对数据编解码
  - 线性激活函数 $z = wx$

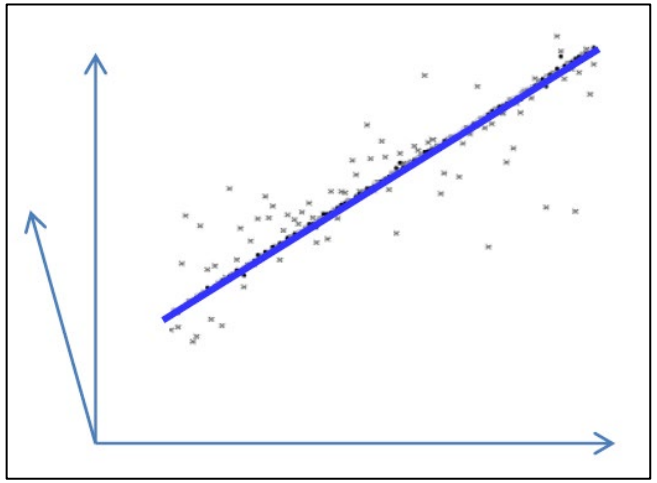
    $\hat{x} = w^T z$
  - 目标函数 $\hat{W} = \underset{W}{\operatorname{argmin}} E[\|x - w^T wx\|^2]$

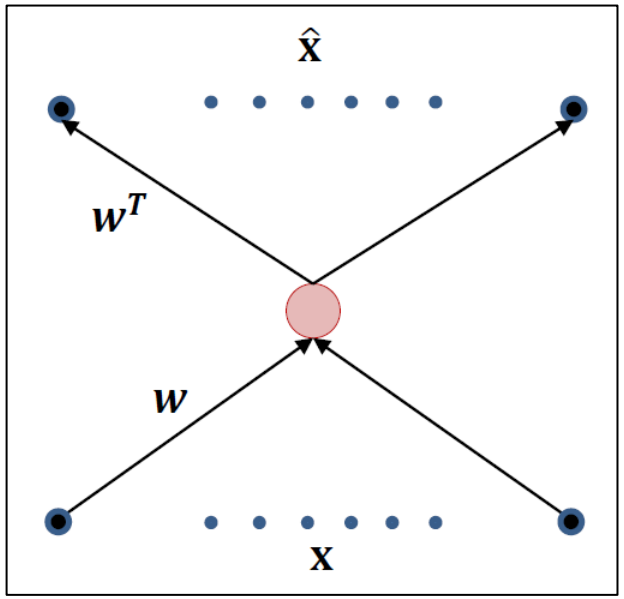


最简单的自动编码器

- AE vs. PCA
  - 编码：寻找第一主成分
  - 解码：将投影点向损失最小方向展开







自动编码器

- 增加至多个隐藏单元
  - 每个线性单元对应一个“方向”
  - 线性激活函数

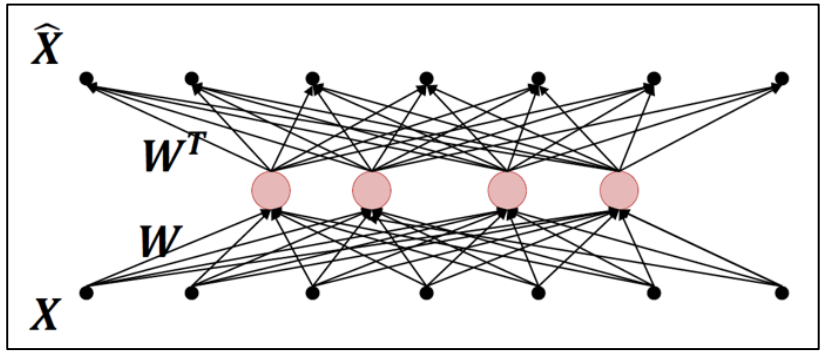
$$z = wx \qquad \hat{x} = w^T z$$

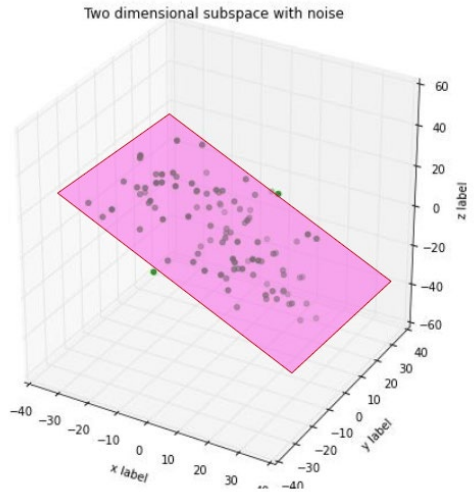
  - 非线性激活函数

$$z = f(x) \qquad \hat{x} = g(z)$$

  - 目标函数

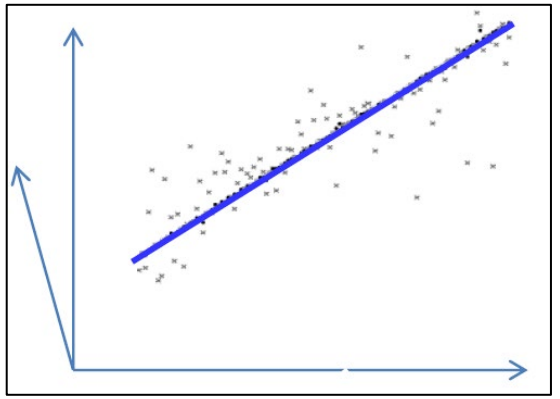
$$\widehat{W} = \text{argmin}_w E\|x - \hat{x}\|^2$$

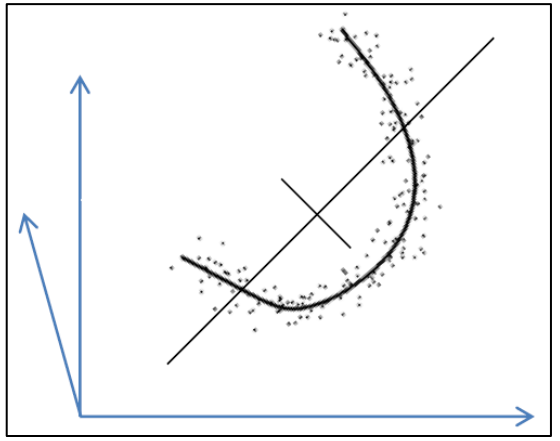




自动编码器

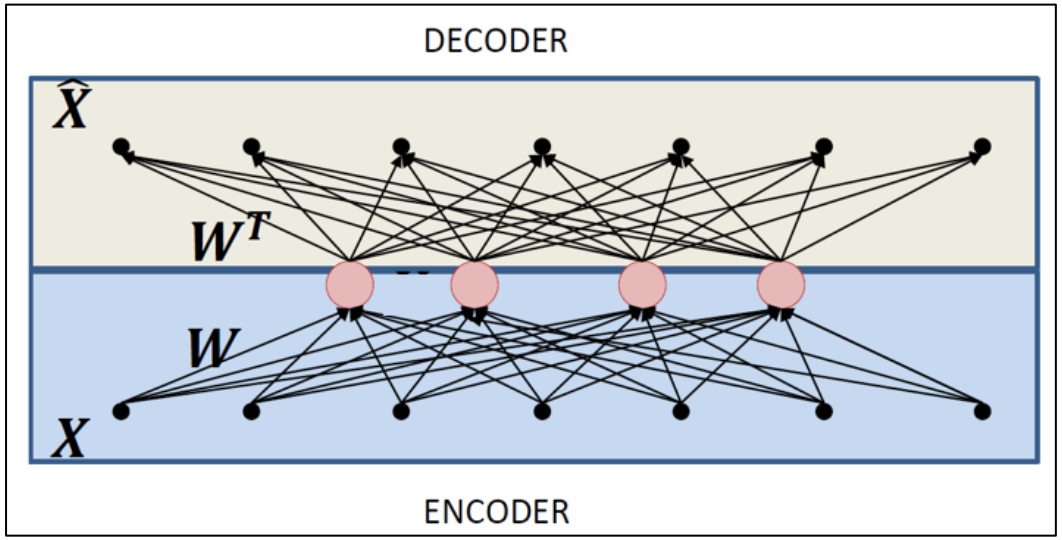
- AE vs. PCA
  - $d(z) < d(x)$
  - 线性激活函数
    - 隐藏层为线性表示
    - 线性PCA
  - 非线性激活函数
    - 隐藏层为非线性表示
    - 核PCA——更强大





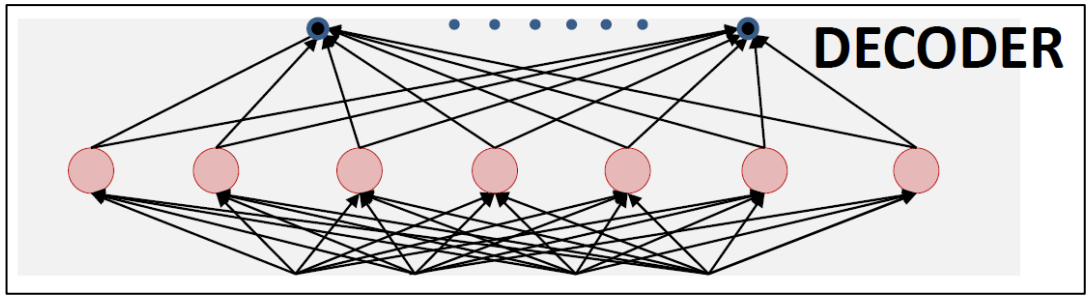
自动编码器

- 基本形式：3层神经网络
  - 输入层
    - 接收输入 x
  - 隐藏层（编码层）
    - 将 x 压缩成隐空间表示 z
  - 输出层（解码层）
    - 通过隐空间表示 z 重构 $\tilde{x}$



自动编码器

- 解码层
  - 只能生成与训练数据非常相似的样本
    - 隐变量 z 来自真实样本经解码器压缩，而非概率意义的分布
- 自编码器并不是真正意义上的“生成模型”

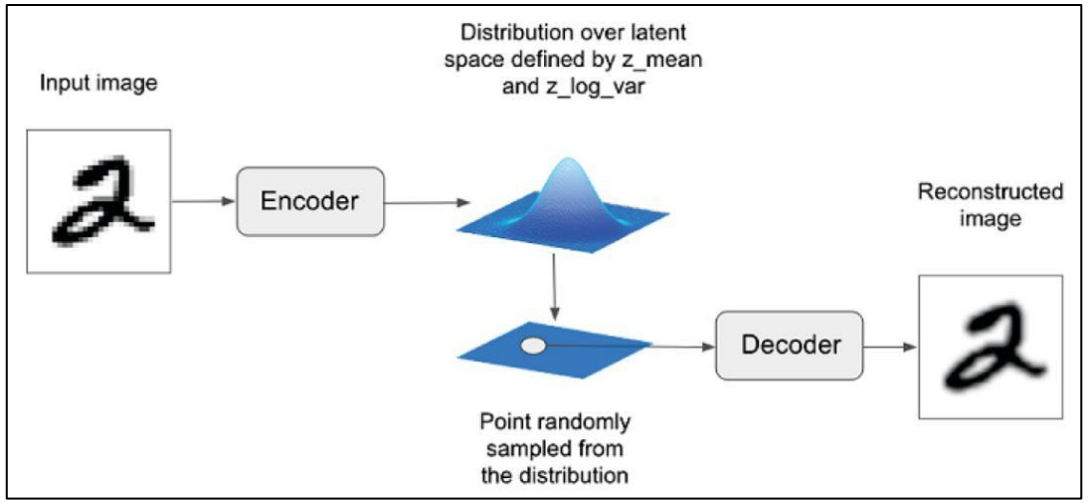


目录

- 隐变量模型
- 自动编码器(Auto-Encoder, AE)
- 变分自编码器(Variational Auto-Encoder, VAE)
- 最新进展

变分自编码器

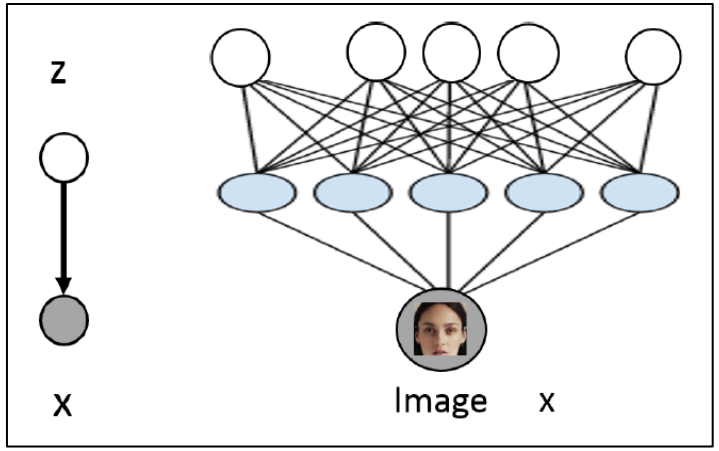
- 变分自编码器：以概率分布刻画隐层，实现新样本采样
  - 约束编码过程，鼓励隐变量大致遵循一个先验分布
  - 通常选择标准正态分布



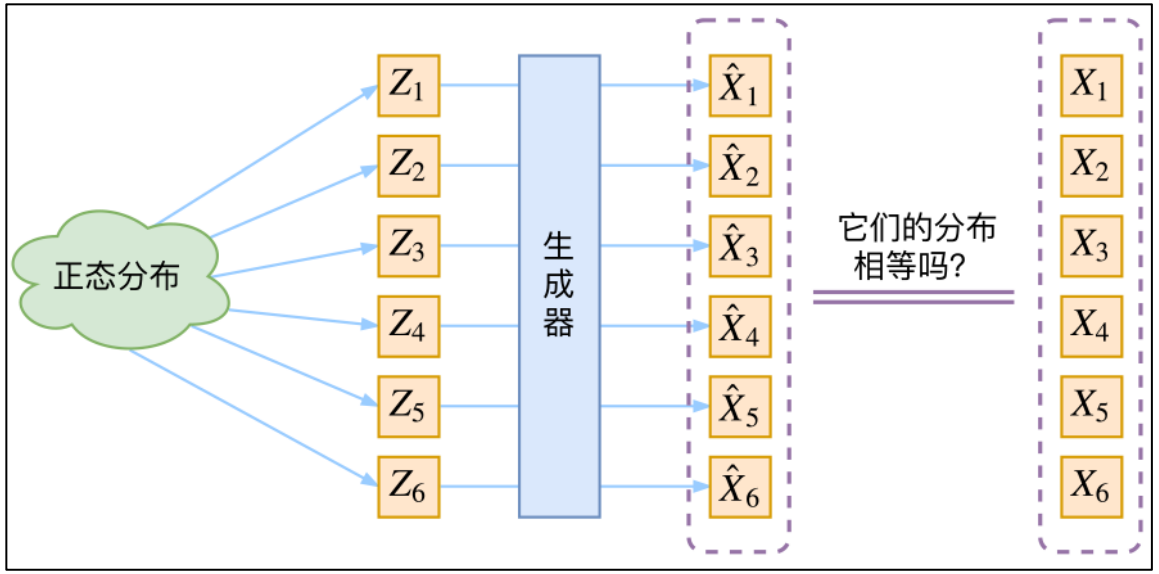
变分自编码器

- 解码器
  - 回顾：目标是从隐变量 z 生成目标数据 x
  - 对于连续数据，有

$$p(x) = \int_z p(z)p(x|z)\,dz$$



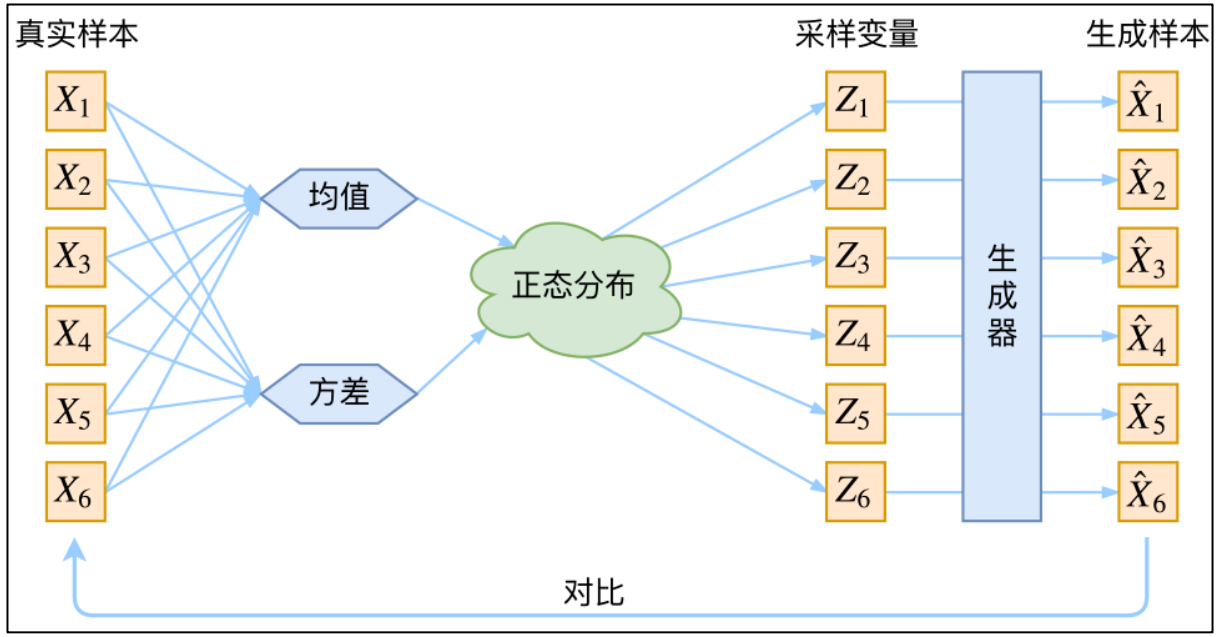
变分自编码器



- 假设 Z 服从特定先验分布，如正态分布
- 目标：训练模型 $X = g(Z)$，将相对简单的隐空间概率分布映射到训练集概率分布

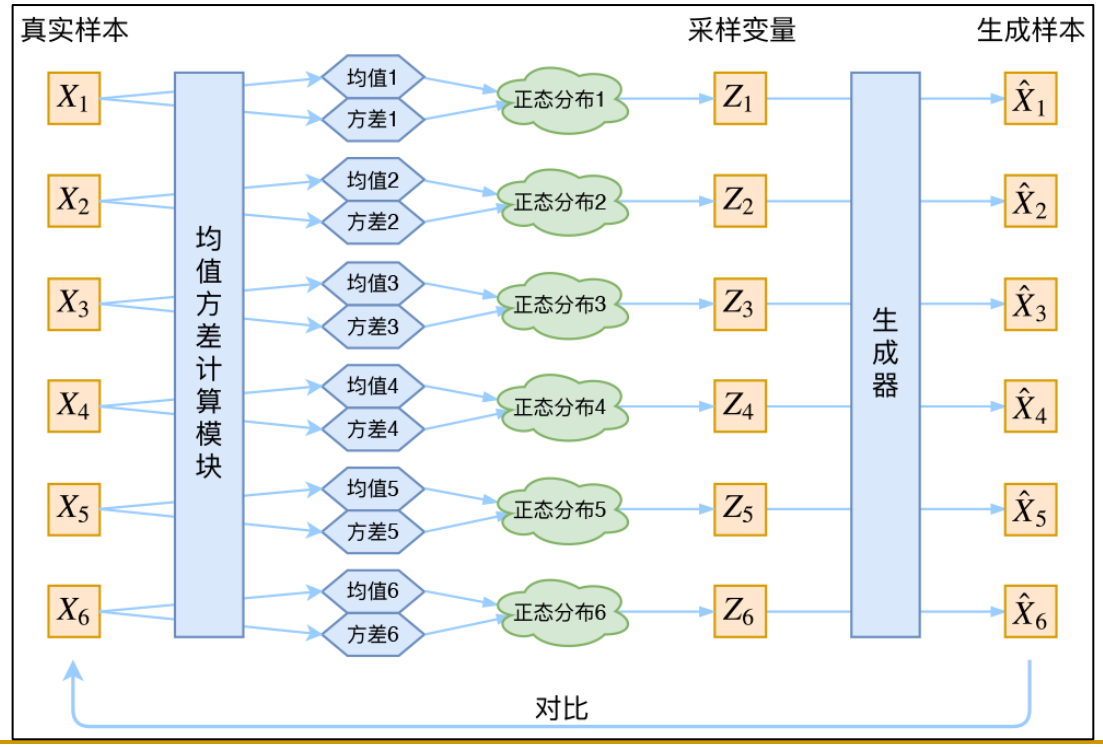
变分自编码器

- 假设$p(z)$是正态分布
  - 具备随机性
  - 重建？无法判断采样的 z_k，是否对应训练样本 x_k



变分自编码器

- 假设 $p(z|x)$ 是正态分布
  - 假设存在一个专属于x_k的分布$p(z|x_k)$



变分自编码器

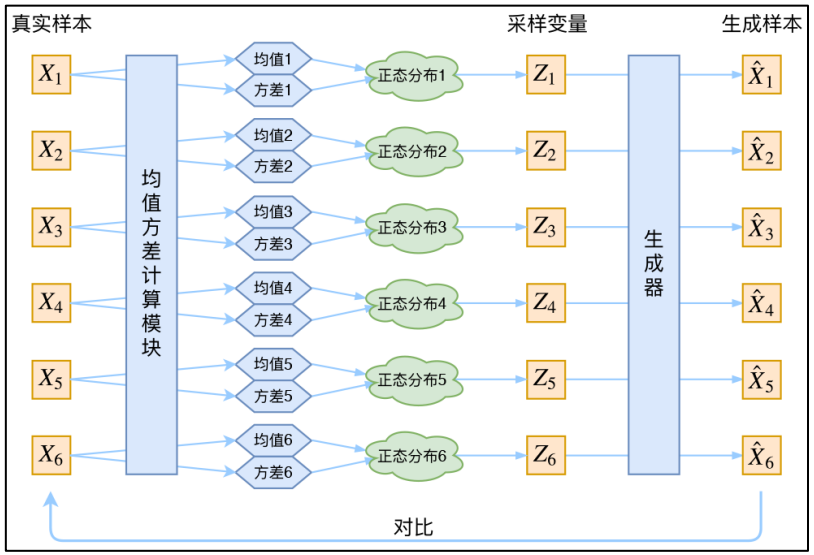
- 构建两个神经网络
  - 计算均值

$$\mu_k = f_1(X_k)$$

  - 计算方差

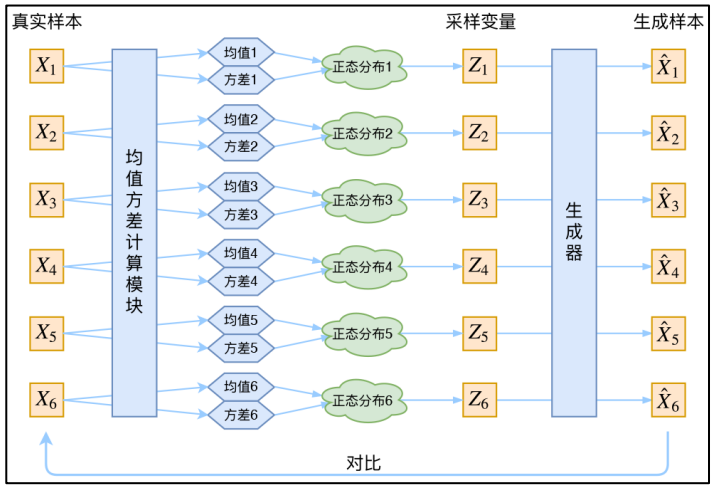
$$\log \sigma_k^2 = f_2(X_k)$$

    - 为什么选择$\log \sigma_k^2$？
      - 确保神经网络输出的方差 σ_k^2 非负



变分自编码器

- 训练：只依赖重构损失？
  - 重构训练数据 x_k 为 $\hat{x}_k$，最小化两者距离
- 问题
  - 降低重构损失 -> 缩小$z_k \sim p(z|x_k)$的误差
  - 此时模型存在“最简单”策略：
    采样$z_k = \mu_k$，此时缺少随机性
  - 即，模型退化成普通自编码器，
    噪声不再起作用



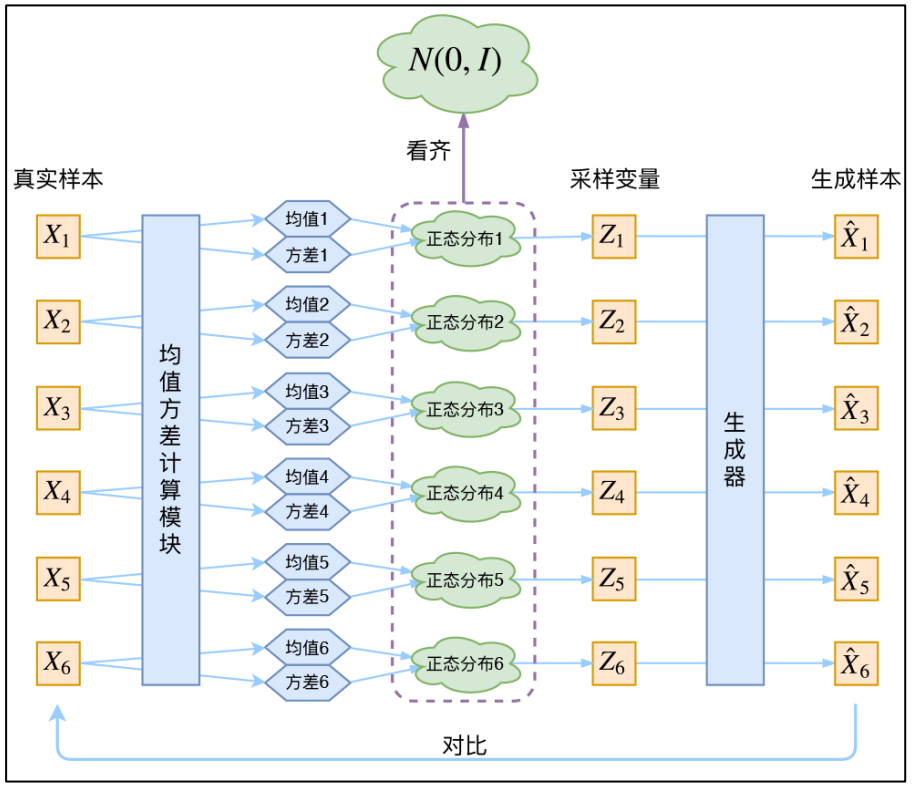
变分自编码器

- 保留随机性：要求 $p(z|x_k)$ 向先验分布对齐

$$
\begin{aligned}
p(Z) &= \sum_x p(Z|X)p(X) \\
&= \sum_x N(0,1)p(X) \\
&= N(0,1)\sum_x p(X) \\
&= N(0,1)
\end{aligned}
$$

  - $p(z|x_k) \neq \mathcal{N}(0,1)$? KL损失

$$
\min D_{KL}(\mathcal{N}(\mu_k, \sigma_k^2), \mathcal{N}(0,1))
$$

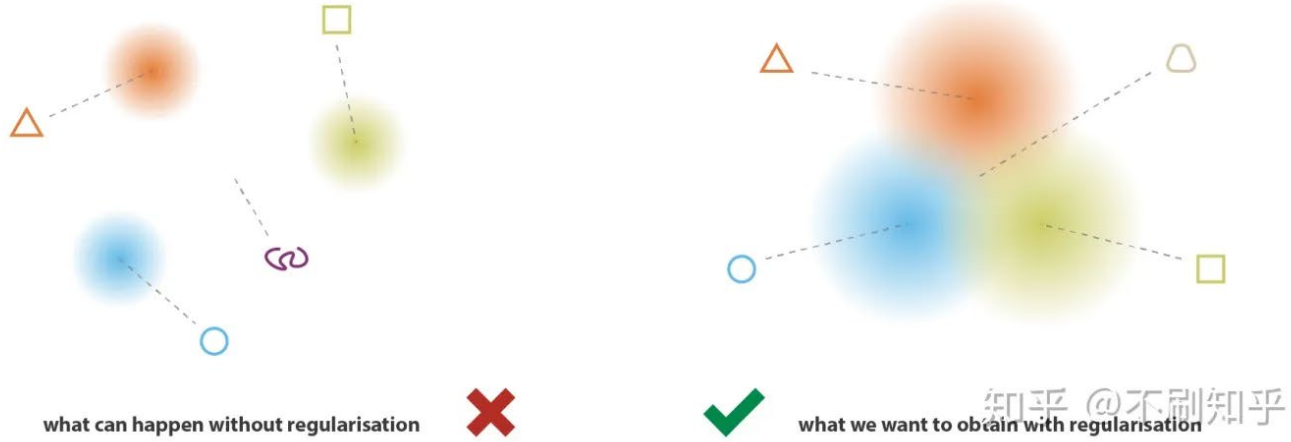


变分自编码器

- 正则化：期望隐空间具有“规律”
  - 连续性（continuity）：隐空间中的相邻点解码后呈现相近内容
  - 完整性（completeness）：针对规定的先验分布，从隐空间采样的点在解码后应提供“有意义”的内容

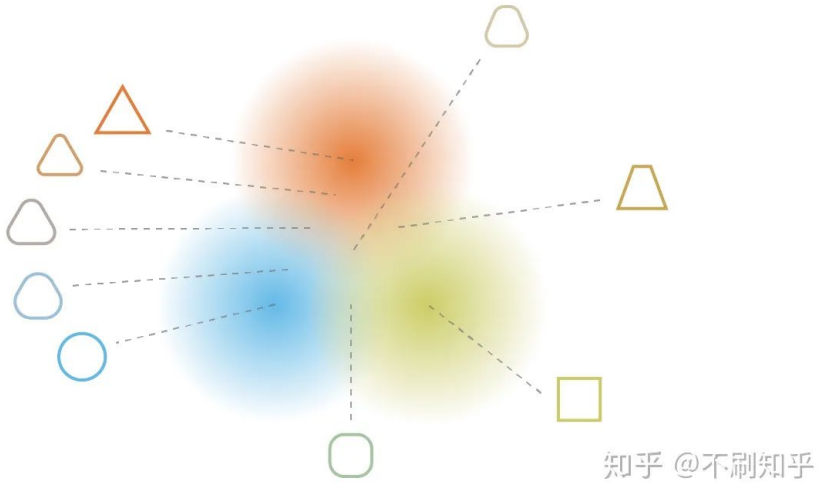
变分自编码器

- 正则化：“产生规律”
  - 要求 $p(z|x_k)$ 与先验分布对齐
    - $\mu = 0$，防止编码分布彼此相距太远，保证了隐空间的连续性
    - $\sigma^2 = 1$，防止出现单点分布，保证了隐空间的完整性



变分自编码器

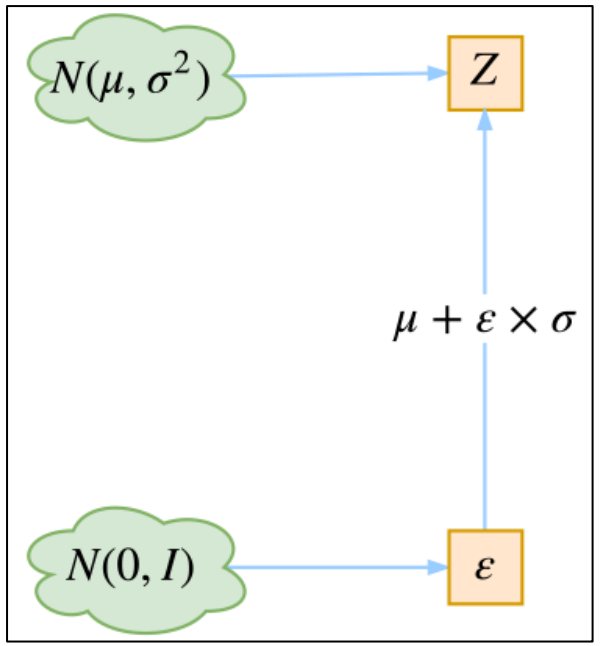
- 正则化：梯度信息
  - 隐空间中，位于来自不同训练数据的两个编码分布的中间点被对应解码为两个分布之间的某个数据



变分自编码器

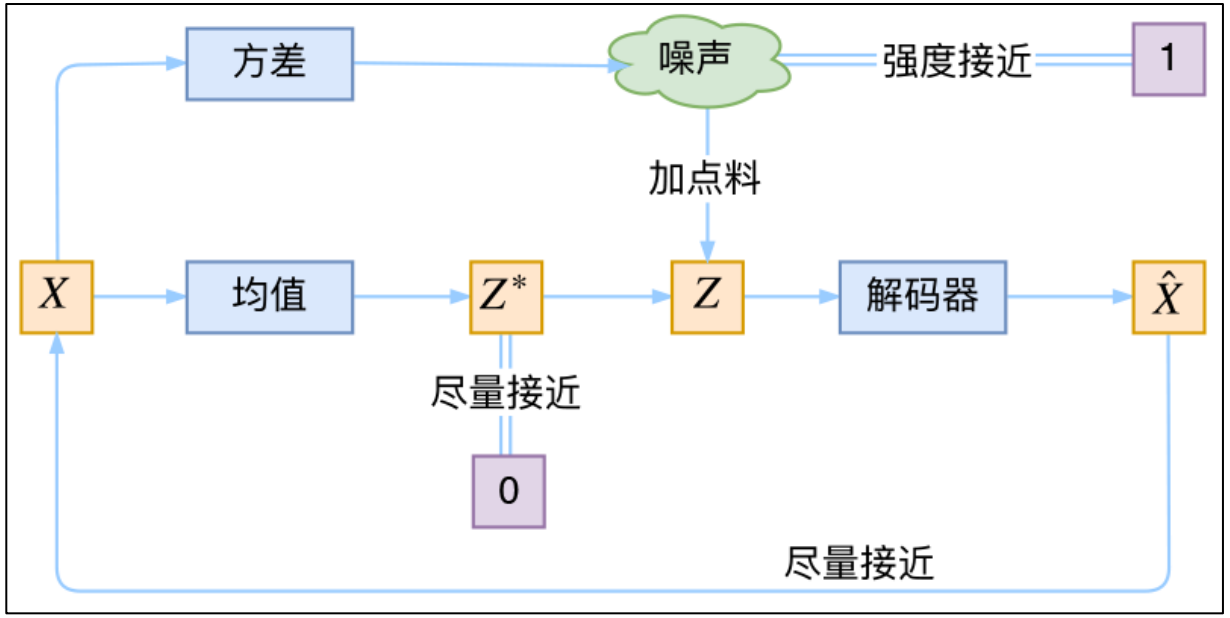
- 如何从隐空间中采样？
  - 随机采样 z 的过程无法求梯度
    - 从 $\mathcal{N}(\mu, \sigma^2)$ 采样
  - 重参数技巧（reparameterization）
    - 从固定的$\mathcal{N}(0,1)$中采样
    - 再施以确定性参数变换

$$z = \mu + \sigma\varepsilon$$



变分自编码器

- 整体结构：兼顾重构准确与隐空间多样
  - 分别编码均值方差，偏置高斯噪声，采样得到 z
  - 解码得到生成样本

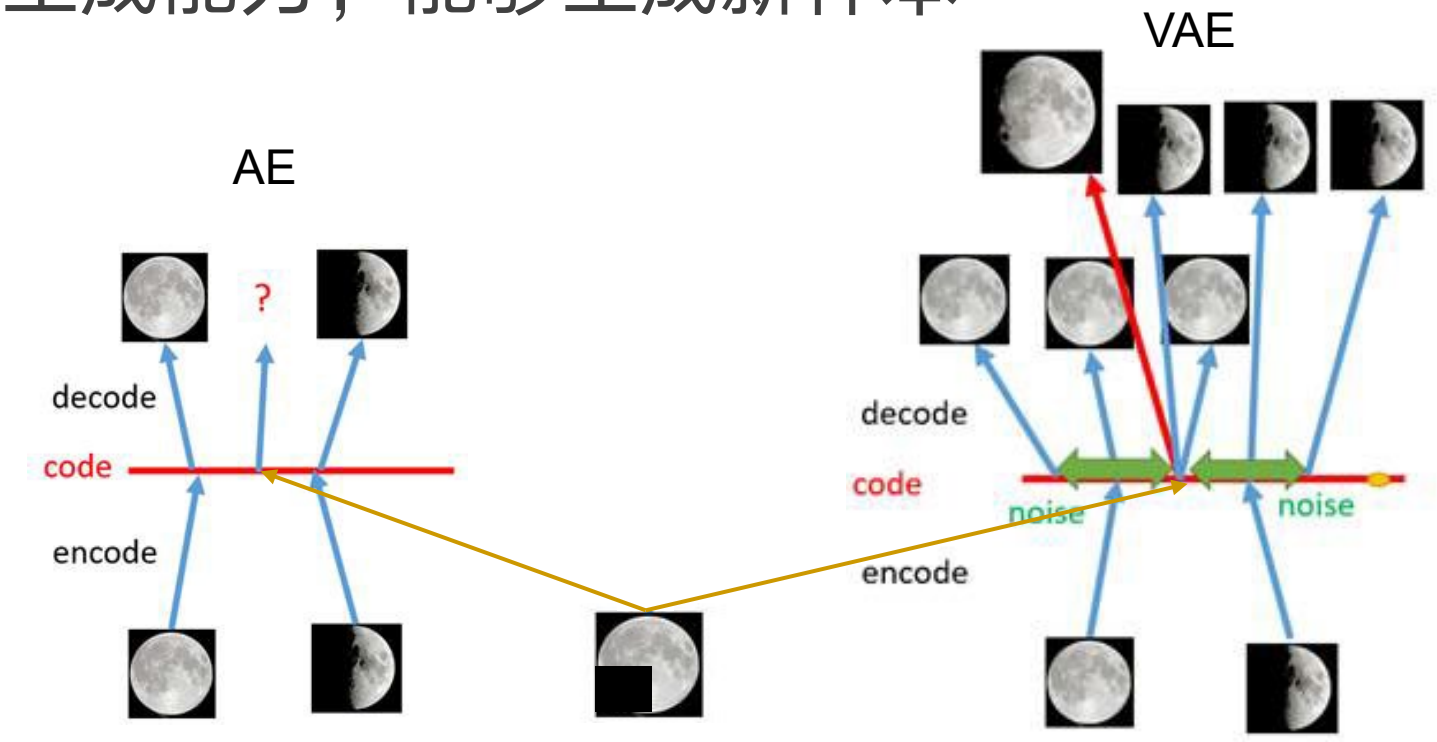


变分自编码器

- 联合优化：动态平衡
  - Decoder未充分收敛：
    - 重构误差 >> KL损失
    - 此时，适当降低噪声（KL损失增加），使拟合更容易（重构误差开始下降）
  - Decoder趋于收敛：
    - 重构误差 < KL损失
    - 噪声增加（KL损失减少），使拟合变得困难（重构误差又开始增加）
    - 此时需提高decoder的生成能力

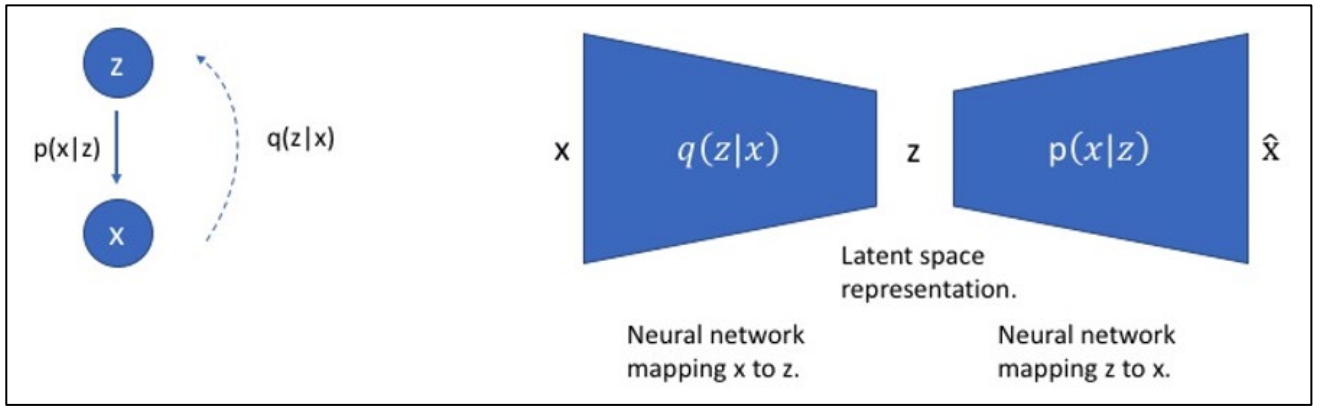
AE vs. VAE

- AE能无损地还原数据集，但不能生成新样本
- VAE具备生成能力，能够生成新样本



变分自编码器

- 实线：解码器 $p(x|z,\theta)$
- 虚线：编码器 $q(z|x,\phi)$



变分自编码器

- 生成模型的学习：似然函数

$$
\begin{aligned}
\log p_{\boldsymbol{\theta}}(\mathbf{x}) &= \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x})\right] \\
&= \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log\left[\frac{p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})}{p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})}\right]\right] \\
&= \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log\left[\frac{p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})}{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\frac{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}{p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})}\right]\right] \\
&= \underbrace{\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log\left[\frac{p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z})}{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\right]\right]}_{=\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x})\ (\text{ELBO})} + \underbrace{\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[\log\left[\frac{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}{p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x})}\right]\right]}_{=D_{KL}(q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})||p_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}))}
\end{aligned}
$$

变分自编码器

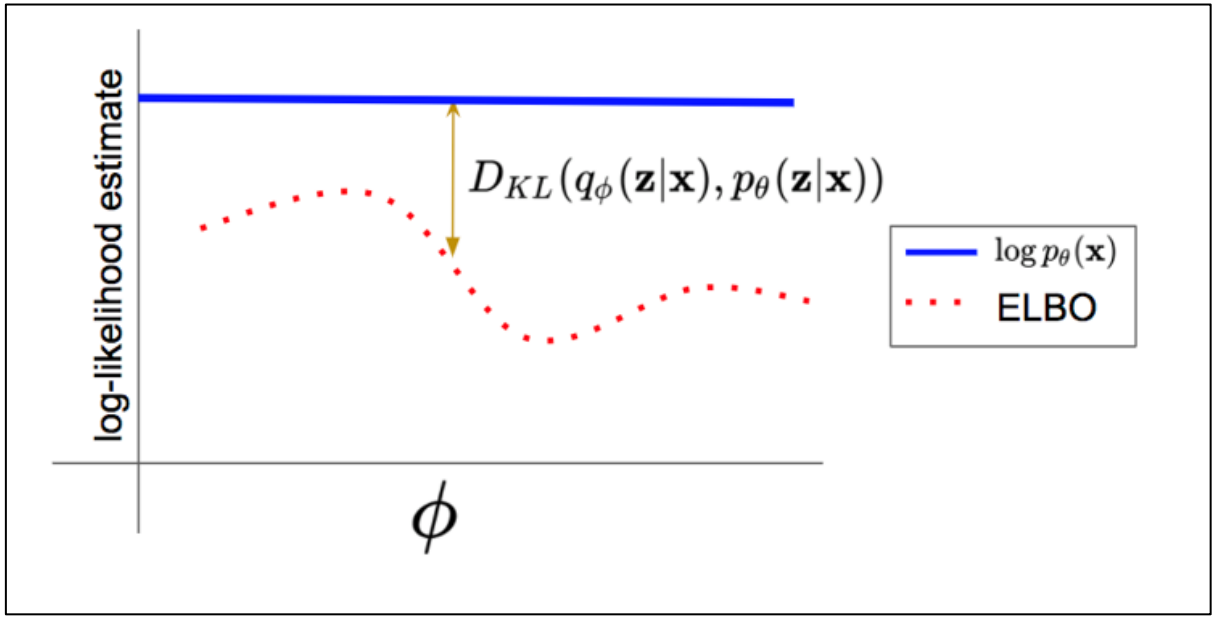
- KL散度 $$D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x})) \geq 0$$
  - 当且仅当$q_\phi(z|x) = p_\theta(z|x)$时，取等号
- 证据下界（ELBO）

$$\begin{aligned}\mathcal{L}_{\theta,\phi}(\mathbf{x}) &= \log p_\theta(\mathbf{x}) - D_{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p_\theta(\mathbf{z}|\mathbf{x})) \\ &\leq \log p_\theta(\mathbf{x})\end{aligned}$$

  - 最大化证据下界
    - 最大化对数似然函数
    - 最小化分布$q_\phi(z|x)$和$p_\theta(z|x)$之间的距离

变分自编码器

- 最大化证据下界
  - 最大化对数似然函数
  - 最小化分布$q_\phi(z|x)$和$p_\theta(z|x)$之间的距离



变分自编码器

- 最大化证据下界

$$L(\theta,\phi;x) = \mathbb{E}_{q_\phi(z|x)} \log \frac{p_\theta(x,z)}{q_\phi(z|x)} = \mathbb{E}_{q_\phi(z|x)} \log \frac{p_\theta(x|z)p_\theta(z)}{q_\phi(z|x)}$$

$$= \mathbb{E}_{q_\phi(z|x)} \log p_\theta(x|z) + \mathbb{E}_{q_\phi(z|x)} \log \frac{p_\theta(z)}{q_\phi(z|x)}$$

$$= \mathbb{E}_{q_\phi(z|x)} \log p_\theta(x|z) - D_{KL}(q_\phi(z|x), p_\theta(z))$$

负交叉熵（负重构损失）

KL损失（正则项）

变分自编码器

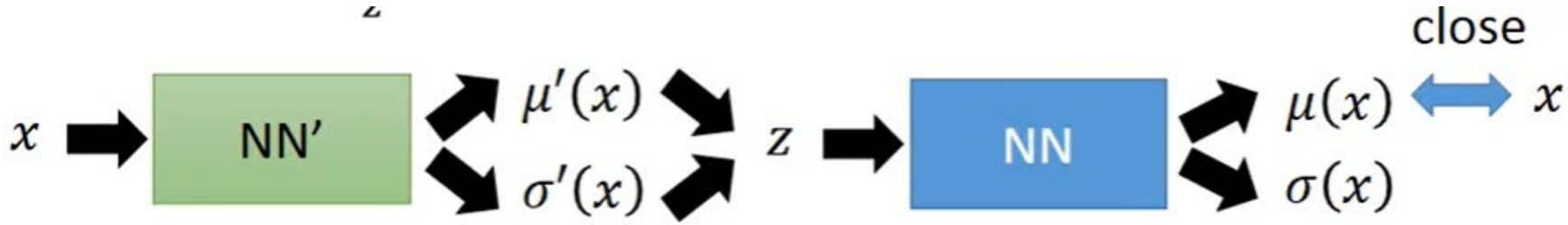
- 最小化KL损失$D_{KL}(q_\phi(z|x), p_\theta(z))$
  - 条件概率$q_\phi(z|x) = \mathcal{N}(\mu, \sigma^2)$
  - 先验概率$p_\theta(z) = \mathcal{N}(0,1)$

$$
\begin{aligned}
&KL\left(N(\mu,\sigma^2)\middle\|N(0,1)\right)\\
&=\int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}\left(\log \frac{e^{-(x-\mu)^2/2\sigma^2}/\sqrt{2\pi\sigma^2}}{e^{-x^2/2}/\sqrt{2\pi}}\right)dx\\
&=\int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}\log\left\{\frac{1}{\sqrt{\sigma^2}}\exp\left\{\frac{1}{2}\left[x^2-(x-\mu)^2/\sigma^2\right]\right\}\right\}dx\\
&=\frac{1}{2}\int \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}\left[-\log\sigma^2+x^2-(x-\mu)^2/\sigma^2\right]dx\\
&=\frac{1}{2}\left(-\log\sigma^2+\mu^2+\sigma^2-1\right)
\end{aligned}
$$

只与**编码器**参数有关

变分自编码器

- 最大化负交叉熵损失$\mathbb{E}_{q_\phi(z|x)} \log p_\theta(x|z)$
  - 在给定$q_\phi(z|x)$（编码器输出）下
    使 $p_\theta(x|z)$（解码器输出）的值尽可能高



变分自编码器

- 求证据下界的梯度
  - 关于解码器参数θ求偏导
    - Monte Carlo近似

$$
\begin{aligned}
\nabla_{\boldsymbol{\theta}} \mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathbf{x}) &= \nabla_{\boldsymbol{\theta}} \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})} \left[\log p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z}) - \log q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\right] \\
&= \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})} \left[\nabla_{\boldsymbol{\theta}} (\log p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z}) - \log q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}))\right] \\
&\simeq \nabla_{\boldsymbol{\theta}} (\log p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z}) - \log q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})) \\
&= \nabla_{\boldsymbol{\theta}} (\log p_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{z}))
\end{aligned}
$$

变分自编码器

- 求证据下界的梯度
  - 关于编码器参数ϕ求偏导

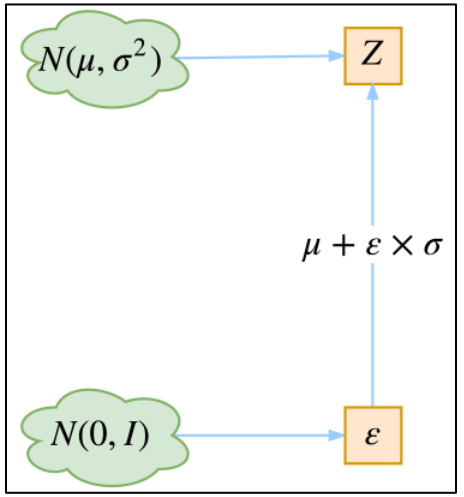
$$\nabla_{\phi}\mathcal{L}_{\theta,\phi}(\mathbf{x}) = \nabla_{\phi}\mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x})}\left[\log p_{\theta}(\mathbf{x},\mathbf{z}) - \log q_{\phi}(\mathbf{z}|\mathbf{x})\right]$$

  - 重参数

$$\mathbf{z} = \mathbf{g}(\boldsymbol{\epsilon}, \boldsymbol{\phi}, \mathbf{x})$$

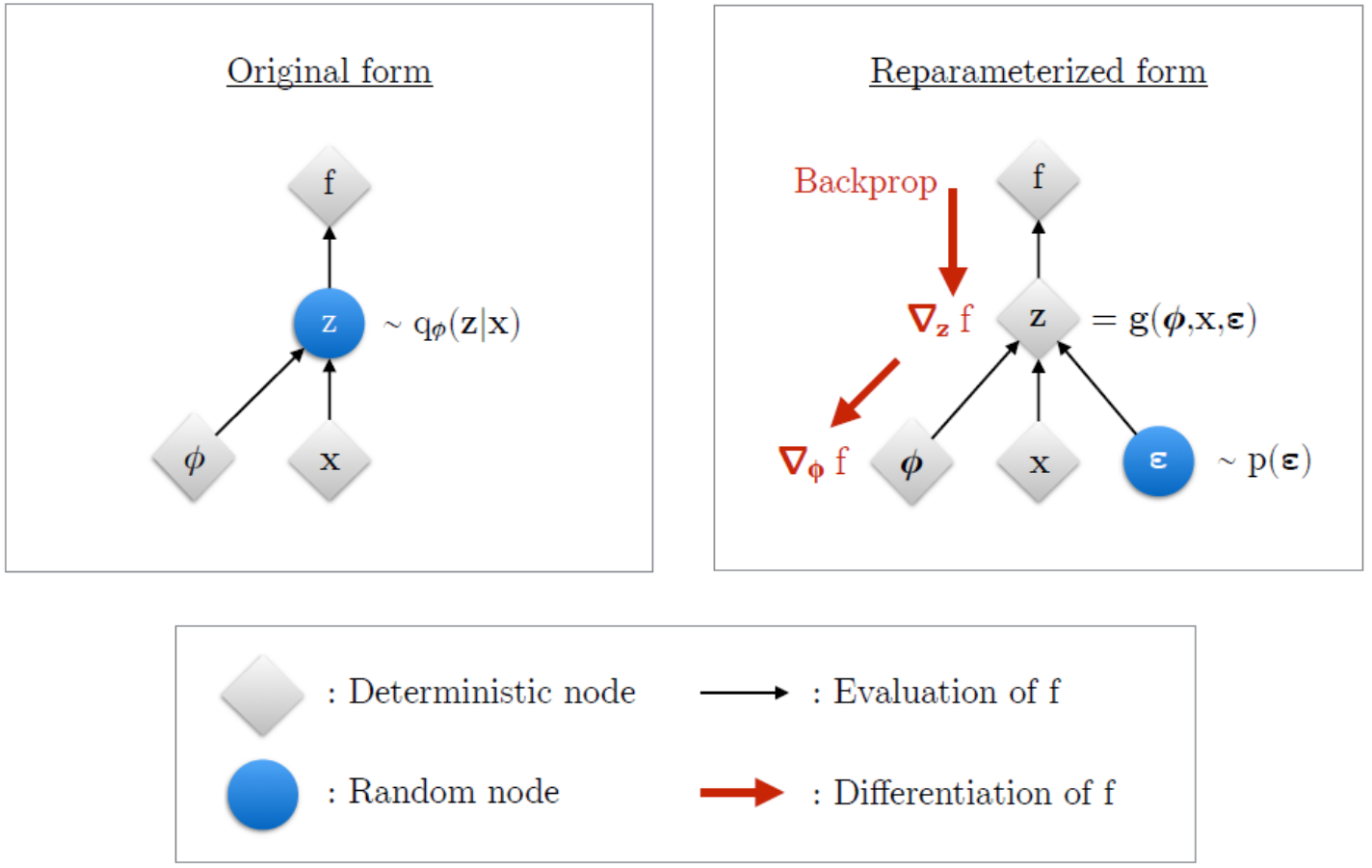
$$\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[f(\mathbf{z})\right] = \mathbb{E}_{p(\boldsymbol{\epsilon})}\left[f(\mathbf{z})\right]$$

$$\begin{aligned}\nabla_{\boldsymbol{\phi}}\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}\left[f(\mathbf{z})\right] &= \nabla_{\boldsymbol{\phi}}\mathbb{E}_{p(\boldsymbol{\epsilon})}\left[f(\mathbf{z})\right] \\ &= \mathbb{E}_{p(\boldsymbol{\epsilon})}\left[\nabla_{\boldsymbol{\phi}} f(\mathbf{z})\right] \\ &\simeq \nabla_{\boldsymbol{\phi}} f(\mathbf{z})\end{aligned}$$



变分自编码器

- 重参数


Original form
f
z
~ $q_\phi(z|x)$
ϕ
x
Reparameterized form
Backprop
f
∇_z f
z
= g(ϕ,x,ε)
∇_ϕ f
ϕ
x
ε
~ p(ε)
: Deterministic node
: Evaluation of f
: Random node
: Differentiation of f

变分自编码器

- 算法

Algorithm 1: Stochastic optimization of the ELBO. Since noise originates from both the minibatch sampling and sampling of $p(\boldsymbol{\epsilon})$, this is a doubly stochastic optimization procedure. We also refer to this procedure as the *Auto-Encoding Variational Bayes* (AEVB) algorithm.

Data:

$\mathcal{D}$: Dataset

$q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$: Inference model

$p_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z})$: Generative model

Result:

$\boldsymbol{\theta}, \boldsymbol{\phi}$: Learned parameters

$(\boldsymbol{\theta}, \boldsymbol{\phi}) \leftarrow$ Initialize parameters

while *SGD not converged* **do**

  $\mathcal{M} \sim \mathcal{D}$ (Random minibatch of data)

  $\boldsymbol{\epsilon} \sim p(\boldsymbol{\epsilon})$ (Random noise for every datapoint in $\mathcal{M}$)

  Compute $\tilde{\mathcal{L}}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathcal{M}, \boldsymbol{\epsilon})$ and its gradients $\nabla_{\boldsymbol{\theta},\boldsymbol{\phi}}\tilde{\mathcal{L}}_{\boldsymbol{\theta},\boldsymbol{\phi}}(\mathcal{M}, \boldsymbol{\epsilon})$

  Update $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ using SGD optimizer

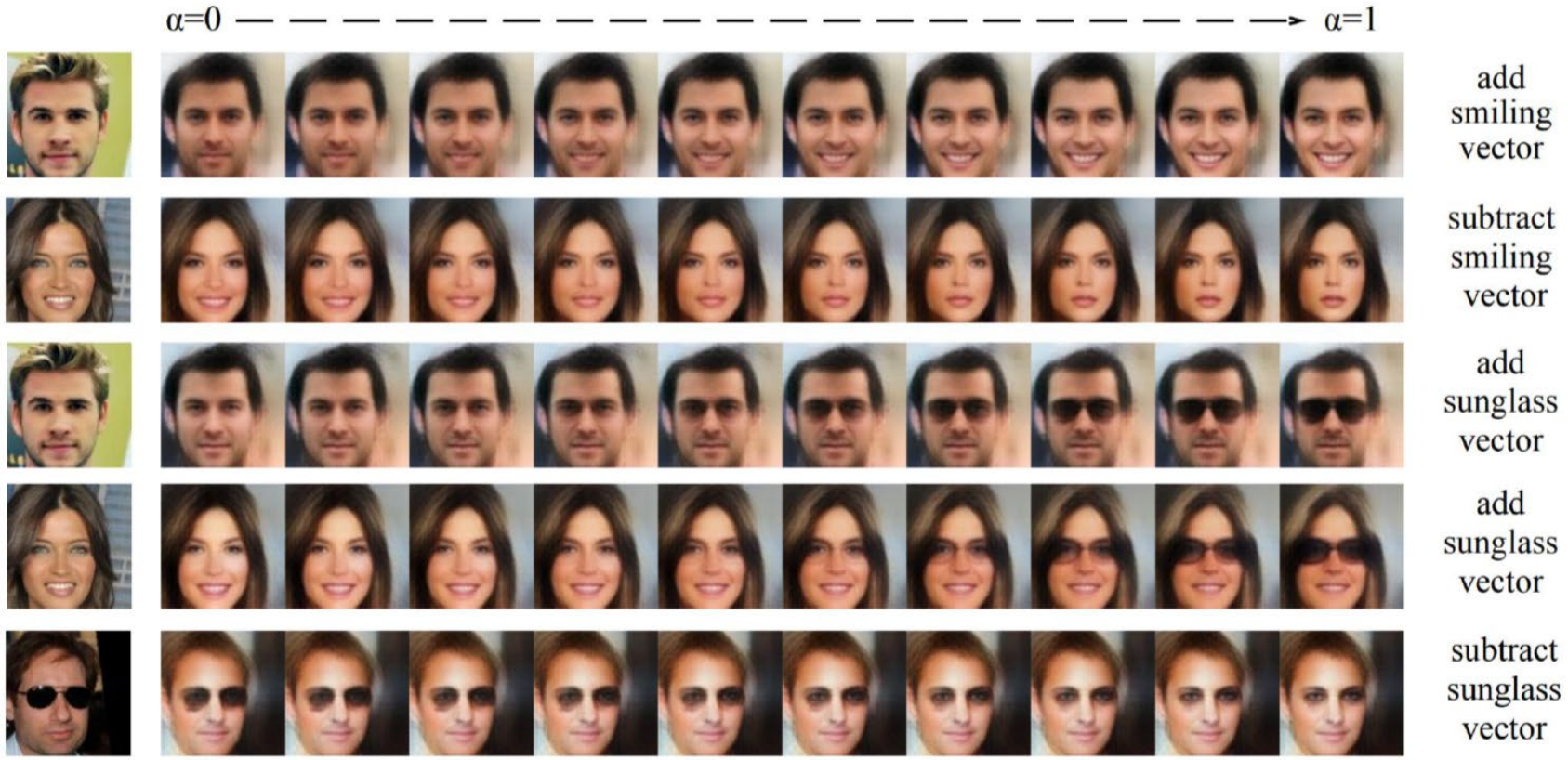
end

变分自编码器

- 条件概率$q_\phi(z|x)$能否选择均匀分布？
  - 如果$q_\phi(z|x)$选择均匀分布，那么$p_\theta(z)$也需要选择为均匀分布
  - 只要二者不相等，就一定存在$q_\phi(z|x) \neq 0$而$p_\theta(z) = 0$的区间
  - 此时KL项发散 $$KL\left(q_\phi(z|x)\,||p_\theta(z)\right) = \int q_\phi(z|x)\log\frac{q_\phi(z|x)}{p_\theta(z)}dz$$

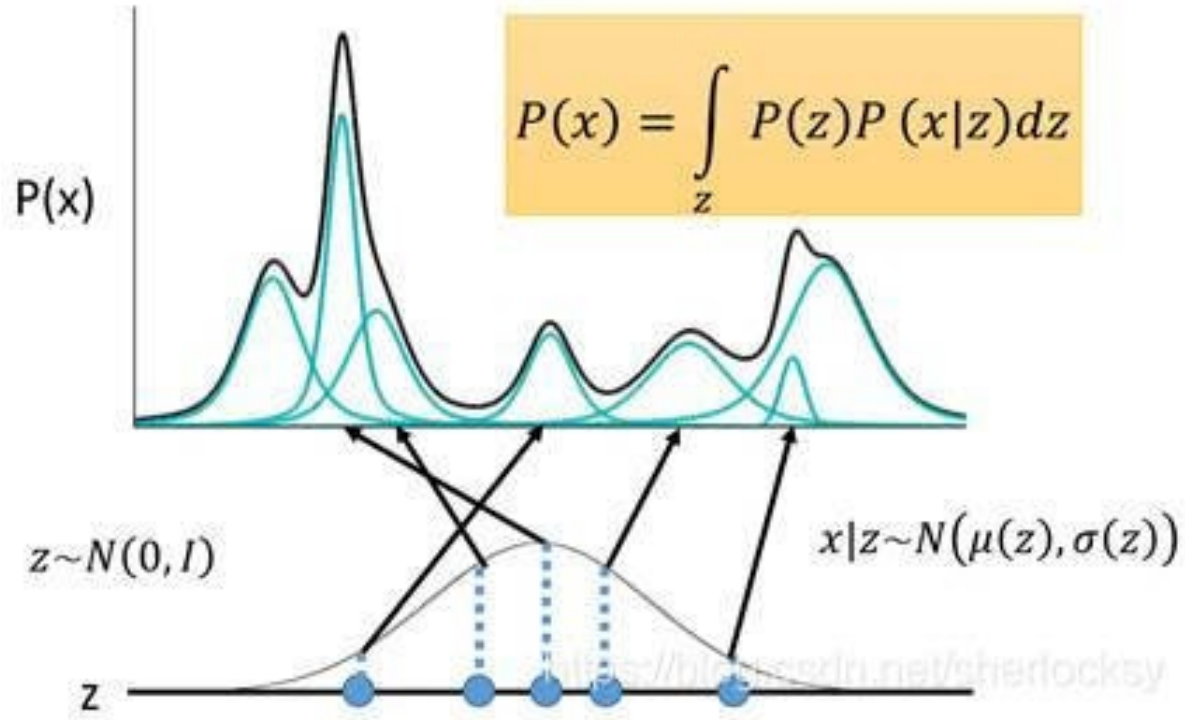
变分自编码器

- 隐空间z的可解释性



变分自编码器 vs. 高斯混合模型

- GMM
  - 有限个高斯分布的混合
  - 线性加权
- VAE
  - 无限个高斯分布的混合
  - 非线性映射

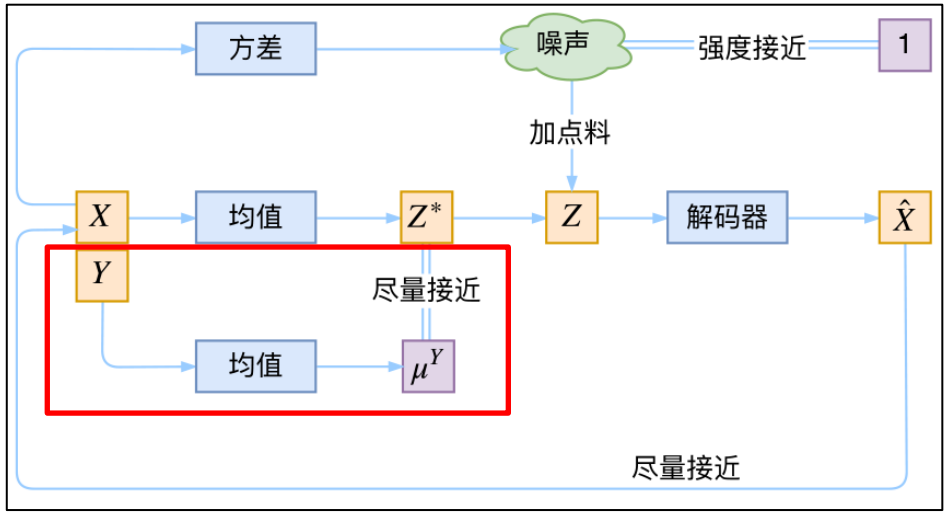


目录

- 隐变量模型
- 自动编码器(Auto-Encoder)
- 变分自编码器(Variational Auto-Encoder)
- 最新进展
  - 条件变分自编码器（conditional VAE）
  - 降噪变分自编码器（denoising VAE ）
  - 对抗自编码器（AAE）

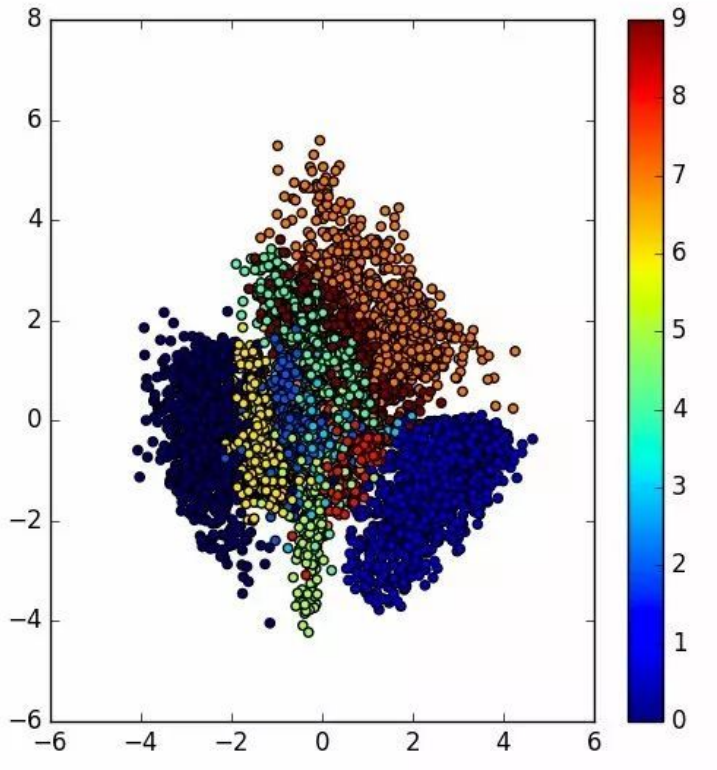
条件变分自编码器

- 条件变分自编码器（CVAE）：生成指定类别/属性的样本
  - 数据 x + 额外信息 y
  - 希望同一类内样本共享隐空间中心（均值） μ^Y
  - 仍保持单位方差



条件变分自编码器：之一

- VAE在隐空间对 z 形成聚类
  - Eg: MNIST数据集
- 半监督学习场景
  - 利用少量标签学习分类，并像VAE生成
  - 判别模型
  - 半监督模型
  - 混合模型



条件变分自编码器：之一

- 判别模型（M1）
  - 使用VAE学习结构化隐空间，利用学到的 z 训练分类器
  - 使用有标签数据
  - 结构和VAE一致

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0}, \mathbf{I}); \qquad p_\theta(\mathbf{x}|\mathbf{z}) = f(\mathbf{x}; \mathbf{z}, \boldsymbol{\theta}),$$

  - 目标函数

$$\log p_\theta(\mathbf{x}) \geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - KL[q_\phi(\mathbf{z}|\mathbf{x})\|p_\theta(\mathbf{z})] = -\mathcal{J}(\mathbf{x})$$

条件变分自编码器：之一

- 半监督模型（M2）
  - 对于无标签数据
    - 将 y 和 z 共同视作隐变量：编码器同时提取隐变量 z 和推断潜在类别 y
  - 编码器：$q_\phi(y,z|x)$
    - 假设 y 和 z 条件独立

$$q_\phi(y,z|x) = q_\phi(y|x)q_\phi(z|x)$$

预测 y：
判别模型分类器

VAE编码器

  - 生成器

$$p(y) = \mathrm{Cat}(y|\boldsymbol{\pi}); \qquad p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\mathbf{0},\mathbf{I}); \qquad p_\theta(\mathbf{x}|y,\mathbf{z}) = f(\mathbf{x};y,\mathbf{z},\boldsymbol{\theta}),$$

条件变分自编码器：之一

$$\log p_\theta(\mathbf{x}) \geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right] - KL[q_\phi(\mathbf{z}|\mathbf{x})\|p_\theta(\mathbf{z})] = -\mathcal{J}(\mathbf{x})$$

- 半监督模型（M2）
  - 似然函数

$$\mathcal{J} = \sum_{(\mathbf{x},y)\sim\widetilde{p}_l} \mathcal{L}(\mathbf{x},y) + \sum_{\mathbf{x}\sim\widetilde{p}_u} \mathcal{U}(\mathbf{x}) + \alpha \cdot \mathbb{E}_{\widetilde{p}_l(\mathbf{x},y)}\left[-\log q_\phi(y|\mathbf{x})\right]$$

有标签数据为判别网络提供了误差

  - 有标签数据

$$\log p_\theta(\mathbf{x},y) \geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x},y)}\left[\log p_\theta(\mathbf{x}|y,\mathbf{z}) + \log p_\theta(y) + \log p(\mathbf{z}) - \log q_\phi(\mathbf{z}|\mathbf{x},y)\right] = -\mathcal{L}(\mathbf{x},y)$$

  - 无标签数据

$$\begin{aligned}\log p_\theta(\mathbf{x}) &\geq \mathbb{E}_{q_\phi(y,\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|y,\mathbf{z}) + \log p_\theta(y) + \log p(\mathbf{z}) - \log q_\phi(y,\mathbf{z}|\mathbf{x})\right] \\ &= \sum_y q_\phi(y|\mathbf{x})(-\mathcal{L}(\mathbf{x},y)) + \mathcal{H}(q_\phi(y|\mathbf{x})) = -\mathcal{U}(\mathbf{x}).\end{aligned}$$

条件变分自编码器：之一

- 混合模型（M1+M2）
  - “两层结构”
    - 使用判别模型学习数据 x 的隐含特征 z_1
    - 对 (z_1, y)而非(x, y)使用半监督生成模型

$$p_\theta(\mathbf{x}, y, \mathbf{z}_1, \mathbf{z}_2) = p(y)p(\mathbf{z}_2)p_\theta(\mathbf{z}_1|y, \mathbf{z}_2)p_\theta(\mathbf{x}|\mathbf{z}_1)$$

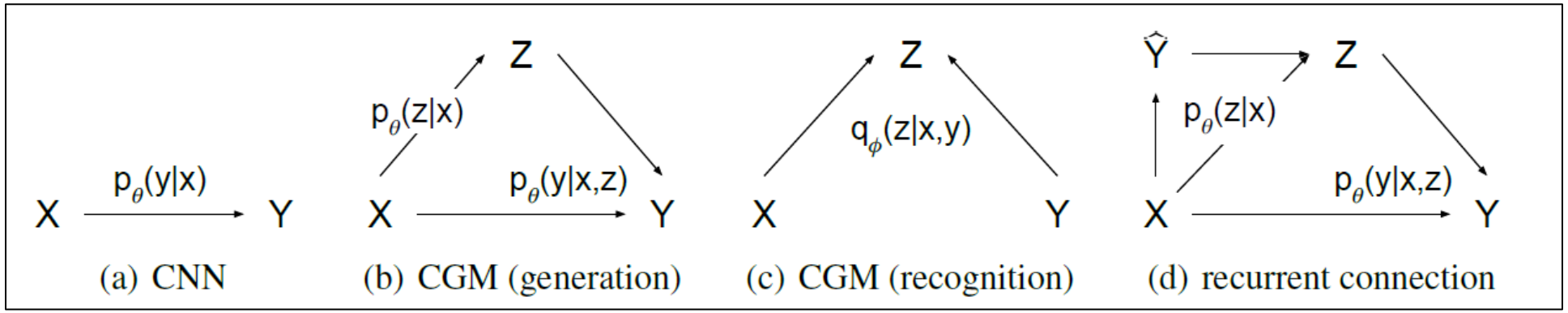
条件变分自编码器：之一

固定类别标签，调整隐空间采样值

(a) Handwriting styles for MNIST obtained by fixing the class label and varying the 2D latent variable **z**

条件变分自编码器：之二

- 在监督学习中，预测标签 y
  - 把 y 看成要生成的数据，x 看成额外信息
  - 先验$p_\theta(z|x)$



条件变分自编码器：之二

- 条件似然函数

$$
\begin{aligned}
\log p_\theta(\mathbf{Y} \mid \mathbf{X}) &= \int_{\mathbf{z}} q_\phi(\mathbf{z} \mid \mathbf{X}, \mathbf{Y}) \log p_\theta(\mathbf{Y} \mid \mathbf{X}) \mathrm{d}\mathbf{z} \\
&= \int_{\mathbf{z}} q_\phi(\mathbf{z} \mid \mathbf{X}, \mathbf{Y}) \log \frac{p_\theta(\mathbf{z}, \mathbf{X}, \mathbf{Y})}{p_\theta(\mathbf{z} \mid \mathbf{X}, \mathbf{Y}) p_\theta(\mathbf{X})} \mathrm{d}\mathbf{z} \\
&= \int_{\mathbf{z}} q_\phi(\mathbf{z} \mid \mathbf{X}, \mathbf{Y}) \log \frac{q_\phi(\mathbf{z} \mid \mathbf{X}, \mathbf{Y})}{p_\theta(\mathbf{z} \mid \mathbf{X}, \mathbf{Y})} \frac{p_\theta(\mathbf{z}, \mathbf{X}, \mathbf{Y})}{q_\phi(\mathbf{z} \mid \mathbf{X}, \mathbf{Y}) p_\theta(\mathbf{X})} \mathrm{d}\mathbf{z} \\
&= \int_{\mathbf{z}} q_\phi(\mathbf{z} \mid \mathbf{X}, \mathbf{Y}) \log \frac{q_\phi(\mathbf{z} \mid \mathbf{X}, \mathbf{Y})}{p_\theta(\mathbf{z} \mid \mathbf{X}, \mathbf{Y})} \mathrm{d}\mathbf{z} + \int_{\mathbf{z}} q_\phi(\mathbf{z} \mid \mathbf{X}, \mathbf{Y}) \log \frac{p_\theta(\mathbf{z}, \mathbf{X}, \mathbf{Y})}{q_\phi(\mathbf{z} \mid \mathbf{X}, \mathbf{Y}) p_\theta(\mathbf{X})} \mathrm{d}\mathbf{z} \\
&= D_{KL}(q_\phi, p_\theta) + \ell(p_\theta, q_\phi)
\end{aligned}
$$

证据下界

条件变分自编码器：之二

- 条件似然函数

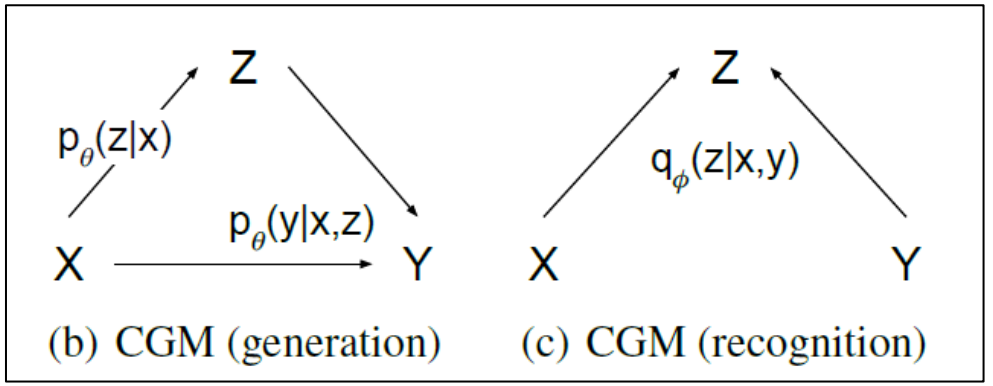
$$
\begin{aligned}
\ell(p_\theta, q_\phi) &= \int_{\mathbf{z}} q_\phi(\mathbf{z} \mid \mathbf{X}, \mathbf{Y}) \log \frac{p_\theta(\mathbf{z}, \mathbf{X}, \mathbf{Y})}{q_\phi(\mathbf{z} \mid \mathbf{X}, \mathbf{Y}) p_\theta(\mathbf{X})} \mathrm{d}\mathbf{z} \\
&= \int_{\mathbf{z}} q_\phi(\mathbf{z} \mid \mathbf{X}, \mathbf{Y}) \log \frac{p_\theta(\mathbf{Y} \mid \mathbf{X}, \mathbf{Z}) p_\theta(\mathbf{Z} \mid \mathbf{X}) p_\theta(\mathbf{X})}{q_\phi(\mathbf{z} \mid \mathbf{X}, \mathbf{Y}) p_\theta(\mathbf{X})} \mathrm{d}\mathbf{z} \\
&= \int_{\mathbf{z}} q_\phi(\mathbf{z} \mid \mathbf{X}, \mathbf{Y}) \log \frac{p_\theta(\mathbf{Z} \mid \mathbf{X})}{q_\phi(\mathbf{z} \mid \mathbf{X}, \mathbf{Y})} \mathrm{d}\mathbf{z} + \int_{\mathbf{z}} q_\phi(\mathbf{z} \mid \mathbf{X}, \mathbf{Y}) \log p_\theta(\mathbf{Y} \mid \mathbf{X}, \mathbf{Z}) \mathrm{d}\mathbf{z} \\
&= -D_{KL}(q_\phi(\mathbf{z} \mid \mathbf{X}, \mathbf{Y}) \mid p_\theta(\mathbf{Z} \mid \mathbf{X})) + \mathbb{E}_{q_\phi}[\log p_\theta(\mathbf{Y} \mid \mathbf{X}, \mathbf{Z})]
\end{aligned}
$$

  - 经验证据下界

$$
\widetilde{\mathcal{L}}_{\mathrm{CVAE}}(\mathbf{x}, \mathbf{y}; \theta, \phi) = -KL\left(q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{y}) \| p_\theta(\mathbf{z}|\mathbf{x})\right) + \frac{1}{L} \sum_{l=1}^{L} \log p_\theta(\mathbf{y}|\mathbf{x}, \mathbf{z}^{(l)})
$$

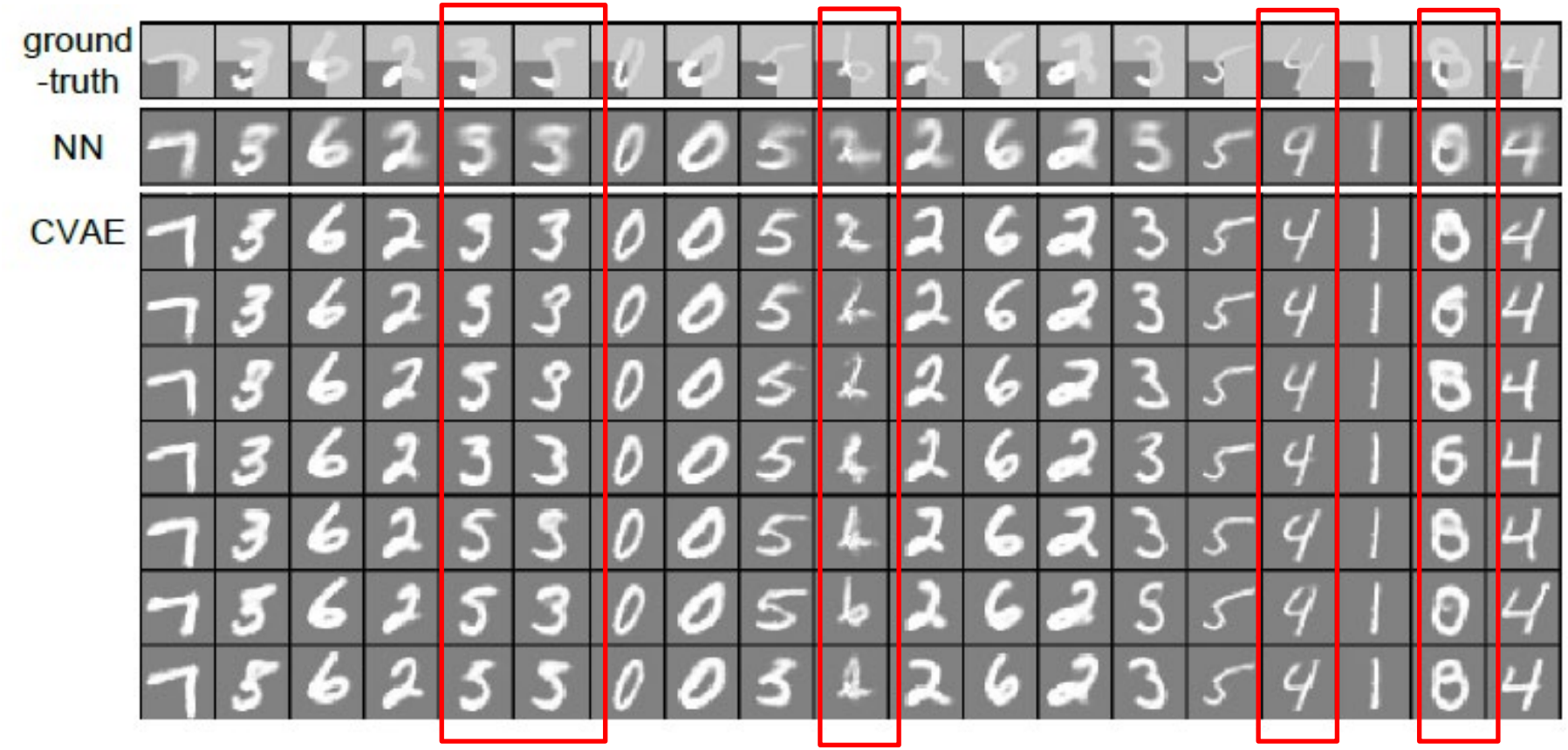
条件变分自编码器：之二

- 网络结构包含三个部分
  - 先验网络$p_\theta(z|x)$
    - 放松z的条件，使得z和x相互独立，即$p_\theta(z|x) = p_\theta(z)$
  - 编码网络$q_\phi(z|x,y)$
  - 解码网络$p_\theta(y|x,z)$
    - 实现one-to-many映射



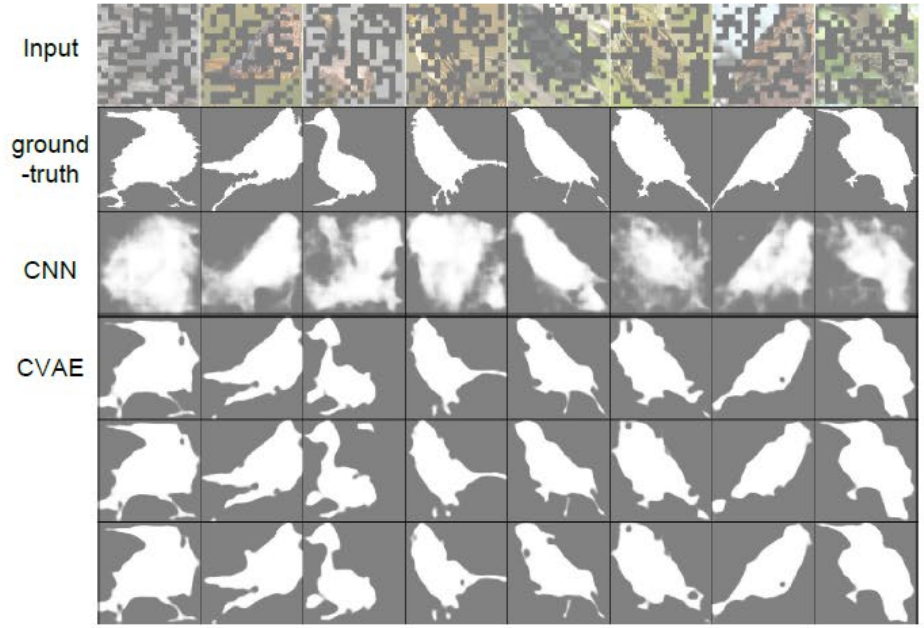
条件变分自编码器：之二

- 预测输入图像中数字的缺失部分



条件变分自编码器：之二

- 预测图片缺失信息



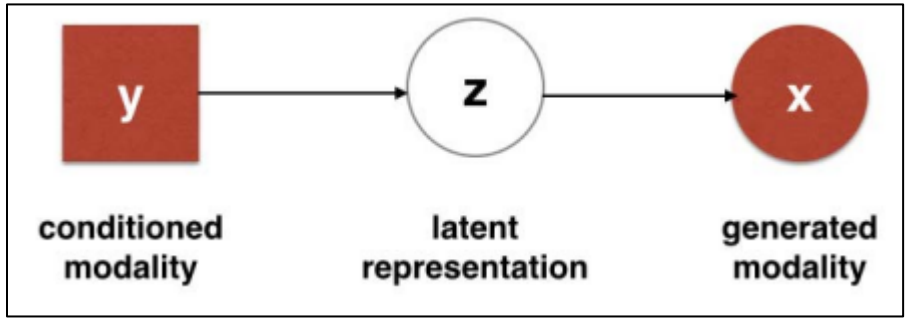
条件变分自编码器：之三

- Conditional Multimodal Auto-Encoder (CMMA)
  - 隐变量 z 由额外信息 y 确定

$$p(\mathbf{x}|\mathbf{y}) = \int p(\mathbf{x}, \mathbf{z}|\mathbf{y})\mathrm{d}\mathbf{z}$$
$$= \int p(\mathbf{x}|\mathbf{z})p(\mathbf{z}|\mathbf{y})\mathrm{d}\mathbf{z}\,.$$

  - x 和 y 条件独立

$$p_\theta(\mathbf{x}|\mathbf{y},\mathbf{z}) = p_\theta(\mathbf{x}|\mathbf{z})$$



条件变分自编码器：之三

- 条件似然函数

$$
\begin{aligned}
&\log p(\mathbf{x}|\mathbf{y}) \\
&= \mathbb{E}_{q(\mathbf{z}|\mathbf{x},\mathbf{y})} \log p(\mathbf{x}|\mathbf{y}) \\
&= \mathbb{E}_{q(\mathbf{z}|\mathbf{x},\mathbf{y})} \log \frac{p(\mathbf{x},\mathbf{z}|\mathbf{y})}{p(\mathbf{z}|\mathbf{x},\mathbf{y})} \\
&= \mathbb{E}_{q(\mathbf{z}|\mathbf{x},\mathbf{y})} \log \frac{q(\mathbf{z}|\mathbf{x},\mathbf{y})}{p(\mathbf{z}|\mathbf{x},\mathbf{y})} + \mathbb{E}_{q(\mathbf{z}|\mathbf{x})} \log \frac{p(\mathbf{x},\mathbf{z}|\mathbf{y})}{q(\mathbf{z}|\mathbf{x},\mathbf{y})} \\
&= \mathrm{KL}\left[q(\mathbf{z}|\mathbf{x},\mathbf{y})||p(z|\mathbf{x},\mathbf{y})\right] + \mathbb{E}_{q(\mathbf{z}|\mathbf{x},\mathbf{y})} \log \frac{p(\mathbf{x},\mathbf{z}|\mathbf{y})}{q(\mathbf{z}|\mathbf{x},\mathbf{y})} \\
&\geq \mathbb{E}_{q(\mathbf{z}|\mathbf{x},\mathbf{y})} \log \frac{p(\mathbf{x},\mathbf{z}|\mathbf{y})}{q(\mathbf{z}|\mathbf{x},\mathbf{y})},
\end{aligned}
$$

条件变分自编码器：之三

- 变分下界

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

$$= \mathbb{E}_{q(\mathbf{z}|\mathbf{x},\mathbf{y})} \log p(\mathbf{x}|\mathbf{z}) + \mathbb{E}_{q(\mathbf{z}|\mathbf{x},\mathbf{y})} \log \frac{p(\mathbf{z}|\mathbf{y})}{q(\mathbf{z}|\mathbf{x}, \mathbf{y})}$$

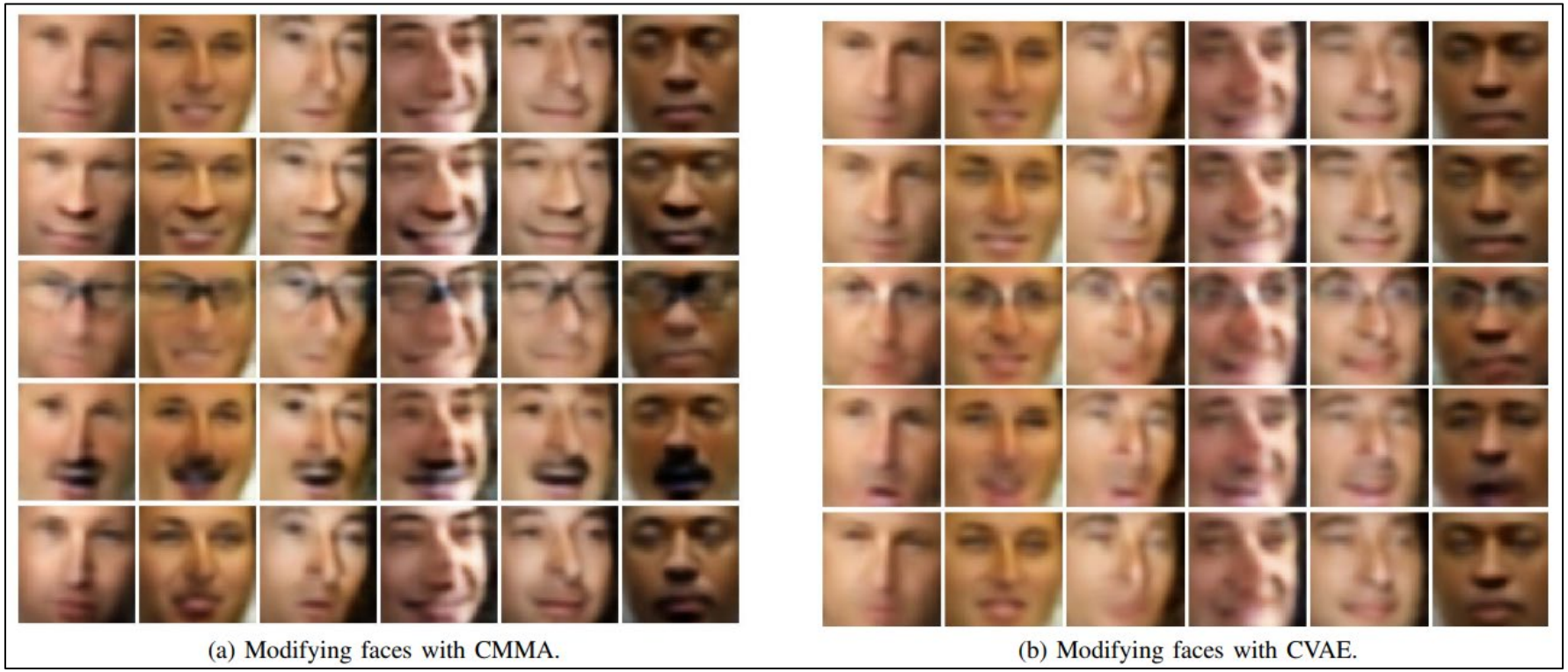
$$= \mathbb{E}_{q(\mathbf{z}|\mathbf{x},\mathbf{y})} \log p(\mathbf{x}|\mathbf{z}) - \mathrm{KL}\left[q(\mathbf{z}|\mathbf{x}, \mathbf{y})||p(\mathbf{z}|\mathbf{y})\right]$$

给定数据x，隐空间表示z和属性y条件独立

Distribution	Parametric form	Representation
$p(\mathbf{z}\|\mathbf{y})$	$\mathcal{N}(f_\mu(\mathbf{y}), e^{f_\sigma(\mathbf{y})})$	$p_f(\mathbf{z}\|\mathbf{y})$
$p(\mathbf{x}\|\mathbf{z})$	$\mathcal{N}(g_\mu(\mathbf{z}), e^{g_\sigma(\mathbf{z})})$	$p_g(\mathbf{x}\|\mathbf{z})$
$q(\mathbf{z}\|\mathbf{x}, \mathbf{y})$	$\mathcal{N}(h_\mu(\mathbf{x}, \mathbf{y}), e^{h_\sigma(\mathbf{x},\mathbf{y})})$	$q_h(\mathbf{z}\|\mathbf{x}, \mathbf{y}) = q_h(\mathbf{z}\|\mathbf{x})$

条件变分自编码器：之三

- 通过修改属性值y来生成数据



(a) Modifying faces with CMMA.

(b) Modifying faces with CVAE.

降噪变分自编码器

- 降噪变分自编码器（Denoising VAE, DVAE）：重构干净样
  - 将输入 x 按分布添加噪声，得到扰动样本 $\tilde{x}$
    输入VAE，并令VAE重构无噪声样本 x
  - 编码器
    - $q_\phi(z|\tilde{x})$

$$q_\phi(\mathbf{z}|\tilde{\mathbf{x}}) = \mathcal{N}(\mathbf{z}|\mu_\phi(\tilde{\mathbf{x}}), \sigma_\phi(\tilde{\mathbf{x}}))$$

      - 高斯分布
    - 原始VAE模型中的编码器$\tilde{q}_\phi(z|x)$

$$\tilde{q}_\phi(\mathbf{z}|\mathbf{x}) = \int q_\phi(\mathbf{z}|\tilde{\mathbf{x}}) p(\tilde{\mathbf{x}}|\mathbf{x}) d\tilde{\mathbf{x}}$$

      - 高斯混合分布
        - 若x离散，则$p(\tilde{x}|x)$取伯努利分布
        - 若x连续，则$p(\tilde{x}|x)$取标准高斯分布

降噪变分自编码器

- 对数似然函数

$$\log p_\theta(\mathbf{x}) \geq E_{\tilde{q}_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\tilde{\mathbf{x}})}\right] = E_{\tilde{q}_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x},\mathbf{z})}{\tilde{q}_\phi(\mathbf{z}|\mathbf{x})}\right] \stackrel{\text{def}}{=} \mathcal{L}_{cvae}$$

- 变分下界

$$\mathcal{L}_{dvae} \stackrel{\text{def}}{=} E_{\tilde{q}_\phi(\mathbf{z}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\tilde{\mathbf{x}})}\right]$$

  - 更紧致下界

$$\log p_\theta(\mathbf{x}) \geq \mathcal{L}_{dvae} \geq \mathcal{L}_{cvae}$$

降噪变分自编码器

- 更紧致下界$\mathcal{L}_{dvae}$

$$
\begin{aligned}
\mathrm{L}_{dvae} &= E_{q'_{\Phi}(z|x)}\left[\log \frac{p_\theta(x,z)}{q_\Phi(z|x')}\right] \\
&= E_{q'_{\Phi}(z|x)}[\log p_\theta(x|z) + \log p(z) - \log q_\Phi(z|x')] \\
&= E_{q'_{\Phi}(z|x)}[\log p_\theta(x|z)] - E_{q'_{\Phi}(z|x)}\left[\log \frac{q_\Phi(z|x')}{p(z)}\right] \\
&= E_{q'_{\Phi}(z|x)}[\log p_\theta(x|z)] - E_{q(x'|x)}E_{q_\Phi(z|x)}\left[\log \frac{q_\Phi(z|x')}{p(z)}\right] \\
&= E_{q'_{\Phi}(z|x)}[\log p_\theta(x|z)] - E_{q(x'|x)}[KL(q_\Phi(z|x'))||p(z)]
\end{aligned}
$$

  - 不能直接表示成KL损失和负交叉熵损失的形式

降噪变分自编码器

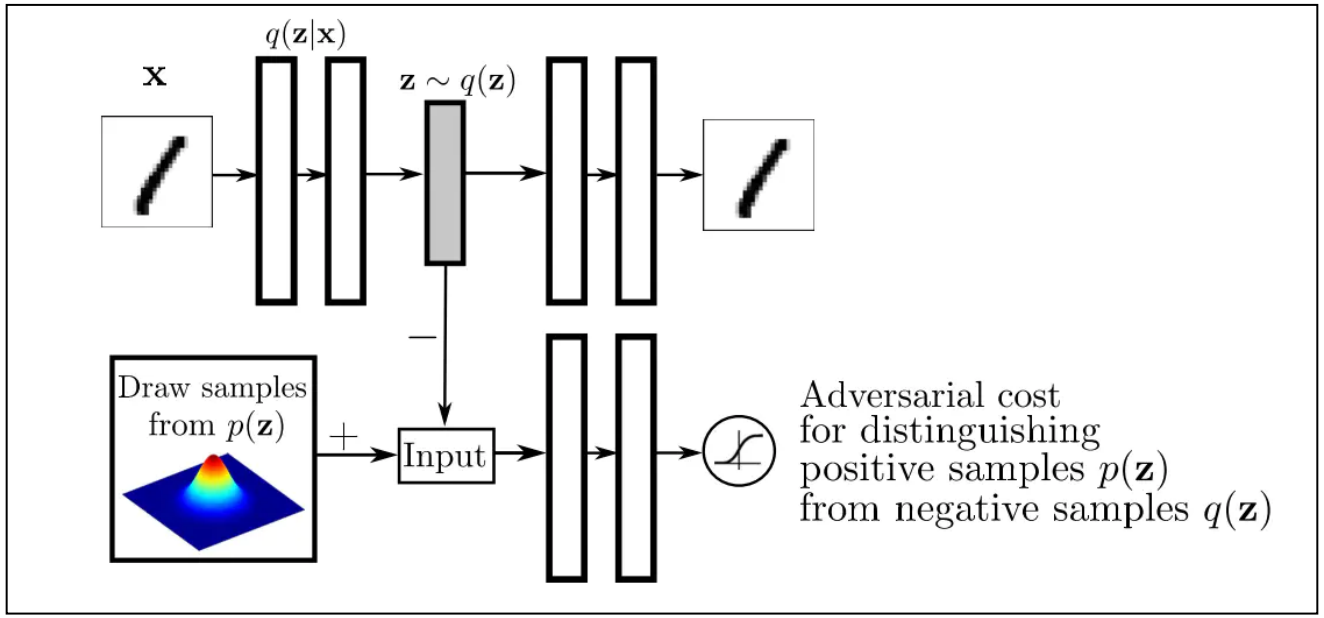
- 求解：蒙特卡洛近似

$$\mathcal{L}_{dvae} = \mathbb{E}_{q(\mathbf{z}|\tilde{\mathbf{x}})}\mathbb{E}_{p(\tilde{\mathbf{x}}|\mathbf{x})}\left[\log \frac{p_\theta(\mathbf{x},\mathbf{z})}{q_\phi(\mathbf{z}|\tilde{\mathbf{x}})}\right] \simeq \frac{1}{MK}\sum_{m=1}^{M}\sum_{k=1}^{K}\log \frac{p_\phi(\mathbf{x},\mathbf{z}^{(k|m)})}{q_\phi(\mathbf{z}^{(k|m)}|\tilde{\mathbf{x}}^{(m)})}$$

- 优点：
  - DVAE能够重构略加破坏的样本，其鲁棒性优于标准VAE
  - $\tilde{q}_\phi(z|x)$是高斯混合模型，更符合真实后验分布
  - DVAE对噪声分布的选择不太敏感，对噪声的强度更敏感

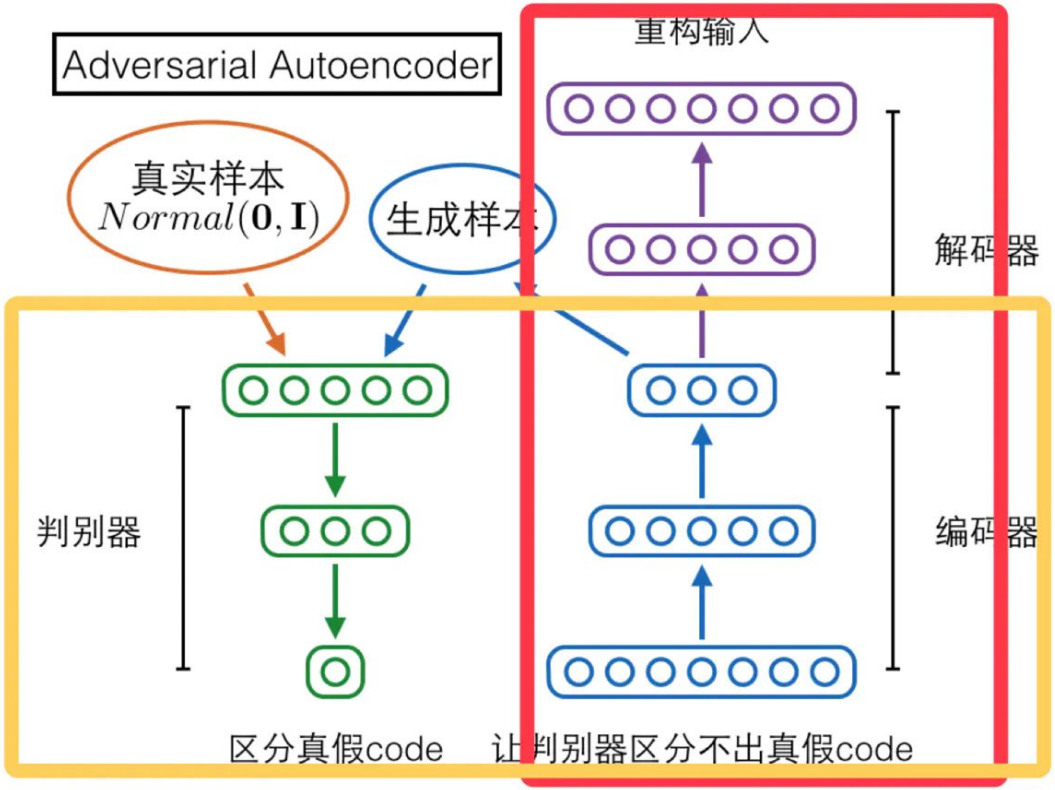
对抗自编码器

- 对抗自编码器（Adversarial AE, AAE）：融入对抗思想
  - 在对抗学习中，编码器同时也是生成器
  - 另外引入一个判别器来判定隐空间特征是否服从想定的数据分布



对抗自编码器

- 主动假设：真实分布为标准正态分布



对抗自编码器

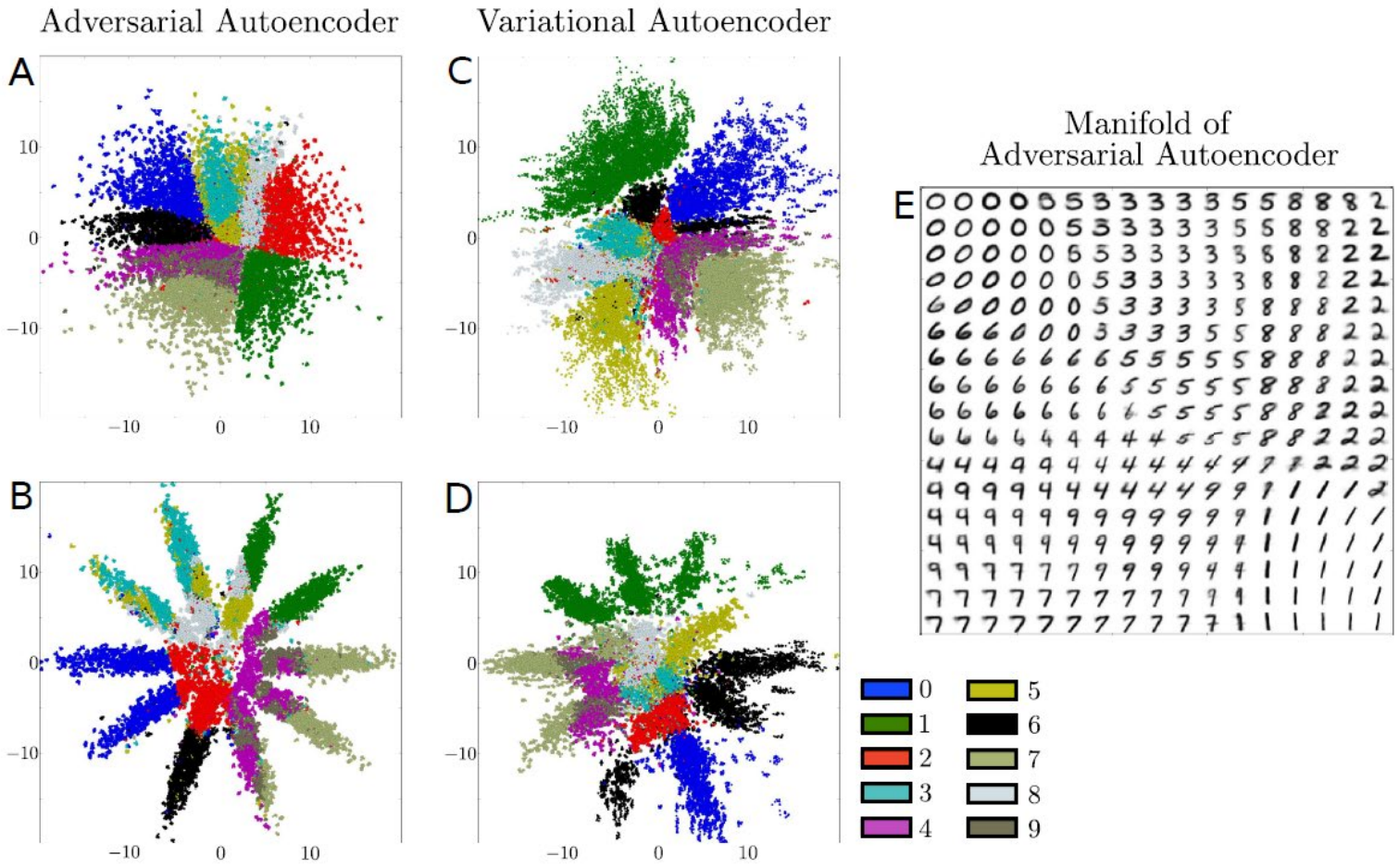
- 模型训练
  - 样本重构阶段：梯度下降更新AE参数、使重构损失函数最小化
  - 正则化约束阶段：交替更新判别网络参数和生成网络（encoder）参数以此提高encoder混淆判别网络的能力
- 生成新数据
  - 直接通过 $p(z)$ 采样所需随机隐变量 z，解码

对抗自编码器

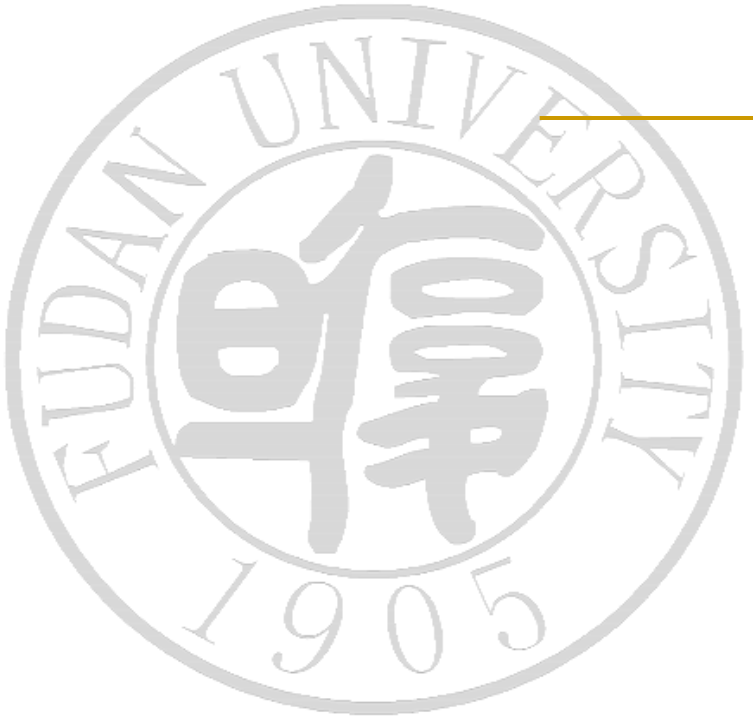
- 与VAE的联系
  - VAE使用KL距离来施加先验分布
  - AAE使用对抗训练来匹配隐编码的聚合后验与先验分布

$$E_{\mathbf{x}\sim p_d(\mathbf{x})}[-\log p(\mathbf{x})] < E_{\mathbf{x}}[\mathrm{E}_{q(\mathbf{z}|\mathbf{x})}[-\log p(\mathrm{x}|\mathrm{z})]] - E_{\mathbf{x}}[\sum_i \log\sigma_i(\mathbf{x}))] + E_{q(\mathbf{z})}[-\log p(\mathbf{z})] + \text{const.}$$
$$= \text{Reconstruction} - \text{Entropy} + \text{CrossEntropy}(q(\mathbf{z}), p(\mathbf{z}))$$

对抗自编码器


Adversarial Autoencoder
Variational Autoencoder
A
B
C
D
E
Manifold of
Adversarial Autoencoder
10
0
−10
0
1
2
3
4
5
6
7
8
9

Thanks!



Questions?