生成模型读书笔记五

2021年1月11日 21:49

1. **VAE**

   有了概率模型和变分推断的知识，现在我们可以介绍VAE

   a. 概率图

   VAE是一种解决推断问题的方法，与前面定义一样，我们有观测变量x，隐变量z，生成分布p(x|z)的参数θ和变分分布q(z|x)的参数φ

   对于这一类问题，我们可以把它的概率图画出来(和概率图那一节的general的图是差不多的，不过这里多了隐变量的参数)

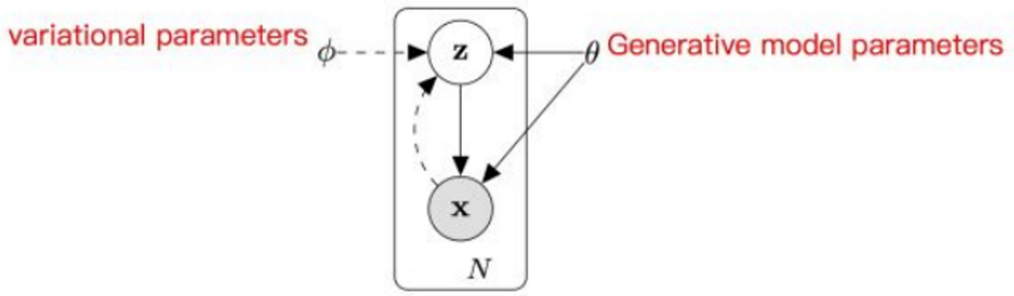


Figure 1: The type of directed graphical model under consideration. Solid lines denote the generative model $p_\theta(\mathbf{z})p_\theta(\mathbf{x}|\mathbf{z})$, dashed lines denote the variational approximation $q_\phi(\mathbf{z}|\mathbf{x})$ to the intractable posterior $p_\theta(\mathbf{z}|\mathbf{x})$. The variational parameters ϕ are learned jointly with the generative model parameters θ.

   b. 生成过程

   在概率模型的推断一节中，我们介绍了生成模型的生成过程，现在来比较正式地描述一个包含隐变量的生成模型的生成过程，以及一些预先假设：

   - 假设：样本间独立同分布
   - 样本的生成过程为：第一步先从一个先验概率分布$P_{\theta^*}(z)$产生z，第二步样本$x^{(i)}$从条件概率分布$P_{\theta^*}(x|z)$中生成
   - $P_{\theta^*}(z)$和$P_{\theta^*}(x|z)$来自假设空间θ，现在就是要解决优化问题找到这个θ*

   c. Variational AutoEncoder(VAE)的提出

   下面来分别解释这两个名词

   i. 变分Variational

   在前面讲解ELBO的一节，我们提到了EM算法可以看作一种应用ELBO的简单算法，它可以解决p(z|x)简单易算的情况，比如GMM，分布形式是简单且确定的，隐变量是离散的，只要固定参数，就可以用$p(x,z)/\sum_z p(x,z)$算出来每种z的后验，那么对于z是高维连续的，p(z|x)难以求解，大量数据的情况，应该怎么做呢？前面也给出了答案，变分法。

   ii. 自编码器AutoEncoder

   AE由一个编码器encoder和一个解码器decoder组成，是一种无监督学习，利用神经网络将输入信息编码到其特征空间，再利用神经网络将这些特征重构为与输入类

似的信息。通常特征z相较输入x的维度要小，只包含重要特征，AE可以通过最小化重构误差来训练。

但是如果我们想从隐空间中生成一个新的样本呢？

AE不是一个生成模型，它无法用于生成，回忆一下前面说过的生成过程，首先从隐变量的先验分布p(z)中采样一个z，再从条件分布p(x|z)中采样一个x。但是AE的隐变量从哪来的呢？从Encoder的输入来的，我们没有学习到z的分布，如果没有encoder的输入，就不能获得一个z，因此为了让模型具有生成功能，就提出了VAE。

iii. VAE原文描述的目的

1) 能够模拟隐空间的生成过程，从而生成新的数据
2) 在给定观测数据x的情况下，能够对隐变量z进行高效的近似后验推断，保持AE representation learning的优秀功能
3) 对x边缘分布的近似，可以应用于一些需要x先验的推断任务，比如图片降噪，补全和超分等。

第一个目的，让它能够模拟p(x|z)p(z)，p(x|z)就是VAE中decoder要做的，第二个目的是学p(z|x)，VAE中的encoder要做的，第三个目的学p(x)=∫p(x|z)p(z)dz, 其中p(z)是我们指定的一个z的先验分布，那到底要怎么学呢？

d. 求解p(z|x)

上述目的的关键就在于如何求解p(z|x) ("关键"可以这么理解：在原来的AE中，通过重构是可以学习到p(x|z)的，p(z)又是事先指定的，所以现在还未解的就剩下p(z|x)了)

这时候就可以用上变分推断了，真实分布p(z|x)比较难学，那就用一个变分分布q(z|x)去近似它，就是要最小化q(z|x)和p(z|x)之间的KL散度，来化简KL

$$
\begin{aligned}
D_{KL}(q(Z|X)||p(Z|X)) &= \mathbb{E}[\log(q(Z|X)) - \log(p(Z|X))] \\
&= \mathbb{E}[\log(q(Z|X)) - \log(\frac{p(X|Z)p(Z)}{p(X)})] \\
&= \mathbb{E}[\log(q(Z|X)) - \log(p(X|Z) - \log(p(Z)))] + \log(p(X)) \\
&= \mathbb{E}[\log(\frac{q(Z|X)}{p(Z)}) - \log(p(X|Z))] + \log(p(X)) \\
&= D_{KL}[q(Z|X)||p(Z)] - \mathbb{E}[\log(p(X|Z))] + \log(p(X))
\end{aligned}
$$

与前面的化简不一样，这里替换p(z|x)的是p(x|z)*p(z) / p(x)，而前面是用联合概率，因为这里我们用神经网络所表示的就是条件分布。

左右项整理一下就得到了ELBO

$$
\log(p(X)) - D_{KL}(q(Z|X)||p(Z|X)) = \mathbb{E}[\log(p(X|Z))] - D_{KL}[q(Z|X)||p(Z)] \text{，}
$$

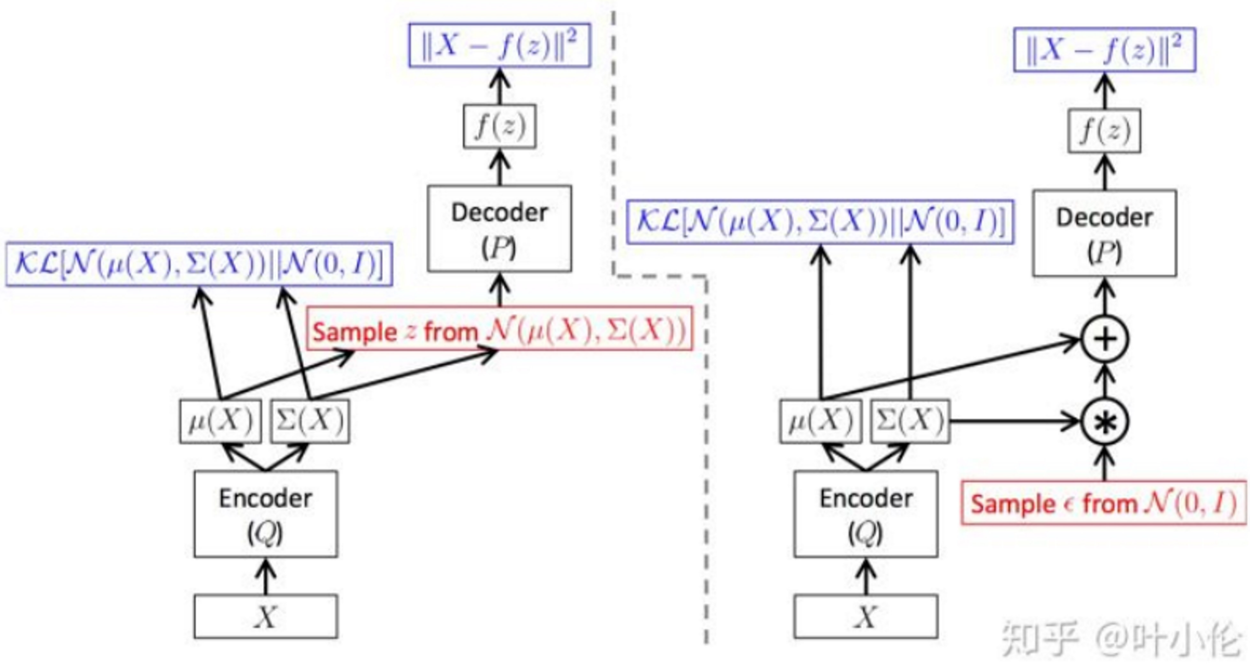
这个化简一下跟上面ELBO的两项分析是一样的，是似然与熵之间的平衡。第一项是似然，希望在给定隐变量z的情况下，能够尽可能地生成观测数据，p(x|z)是可以从decoder算出来的，第二项是z的后验分布和真实先验分布的距离，先验分布p(z)是我们预先定义的，一般会选择一个熵较大(变化较多，覆盖面较大)的一个分布，q(z|x)是从encoder出来的，所以定义好了p(z)之后第二项也可以算， 因此应用变分推断的最大化ELBO即最小化KL距离我们就可以通过最大化上面的ELBO，求得近似分布q。

e. 计算ELBO梯度

上面的分析解决了p(z|x)可算的问题，但是还有一个问题是，因为KL项要计算的是两个分布之间的距离，那么z就应该是连续的才能表示分布，而不是像AE一样直接从encoder生成离散的z，因此，这里encoder的输出是z分布的参数。在VAE中假设z先验分布p(z)是高斯，因此encoder的输出是分布的参数均值μ和方差σ。那么从学习的分布获取z就需要采样，如果直接从分布中采样，采样这个操作就在梯度传递的路径上了，无法求梯度，所以这里使用了前面ELBO求梯度时的第二个方法，重参数法。
(无法求梯度怎么理解：求梯度，肯定得有个函数的形式，如z=f(x)才能够求，现在想想看我们整个过程是怎么样的，从x开始，μ=En1(x), σ=En2(x)，z~N(μ，σ)，x' = De(z), Loss=f(x',x)，从loss开始反向传播求梯度，到了z~N(μ，σ)这一步要怎么算呢，没法算，从而破坏了梯度传递的连续性，无法传播梯度。)
因此用重参数法。任何一个高斯分布，都可以通过z=μ+σ*e，e~N(0,1)得到，替换直接从encoder学习的分布中采样。首先从标准正态分布中采样e，再通过线性变换来得到z，这时候采样操作就在求梯度的路径以外了，可以进行梯度计算。

$$\epsilon \sim q(\epsilon)$$
$$\mathbf{z} = \mathbf{z}(\epsilon\,;\ \lambda),$$

直接采样和重参数采样的两个过程如下面示意图所示：



可见，右边的梯度反向传播路径不需要经过采样步骤。

f. VAE的具体过程和loss

上面推导到了ELBO，我们知道目标就是优化ELBO，但还没有把VAE的具体loss函数给写出来，这一步继续把具体的分布放进去，看看具体的计算式子是什么。
先来捋一捋VAE的整个流程，可以参照上图

1) VAE由encoder和decoder组成，encoder负责从输入空间到隐空间，即q(z|x)，因为z要连续，因此encoder 的输出是z分布的参数，decoder负责从隐空间到重构的出入空间，即p(x|z)，目标是要优化ELBO = E[logp(x|z)] + KL[q(z|x)||p(z)]

2) 根据先验知识预先定义计算中需要的分布：看计算ELBO要用到的分布，p(x|z)，网络直接输出x，不需要特别规定；q(z|x)，定义为高斯分布，因此encoder输出的参数是均值μ(x)和方差σ(x)；p(z)，定义为标准正态分布

接上一节的这一步开始

$$\mathbb{E}[\log(p(X|Z))] - D_{KL}[q(Z|X)||p(Z)]$$

第一项是条件似然，希望数据集的样本x在p(x|z)这个分布中的概率尽量大，换个方向想，也就是说，p(x|z)指从隐空间z生成x，而这些x应该就是输入数据集中的样本，也就是希望能够重构数据集中的样本，所以第一项可以用重构误差计算，对每一个输入的样本点xi都计算重构误差, de(z)表示解码器decoder的输出

$$E[\log p(x|z)] = ||x - de(z)||^2$$

第二项是KL散度，如前所述，我们将z的先验p(z)定义为标准正态分布, q(z|x)也是高斯分布，它的参数是μ(x)和σ(x)，所以把这些具体的分布代进去计算可得

$$KL[q(z|x) \,||\, p(z)] = KL[N(\mu(x), \sigma(x)) \,||\, N(0,1)]$$

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

结果分为三项积分，第一项是∫-logσ²N(μ，σ)dx，里面N就是一个高斯分布，积分自然等于1，因此这项等于-logσ²；

第二项，∫x²N(μ，σ)dx = $E_{x\sim N(\mu,\ \sigma)}[x^2]$，这是高斯分布N的二阶矩，高斯分布的二阶矩是μ²+σ²；第三项，-(x-μ)²/ σ²，就是-方差除以方差，因此是-1。

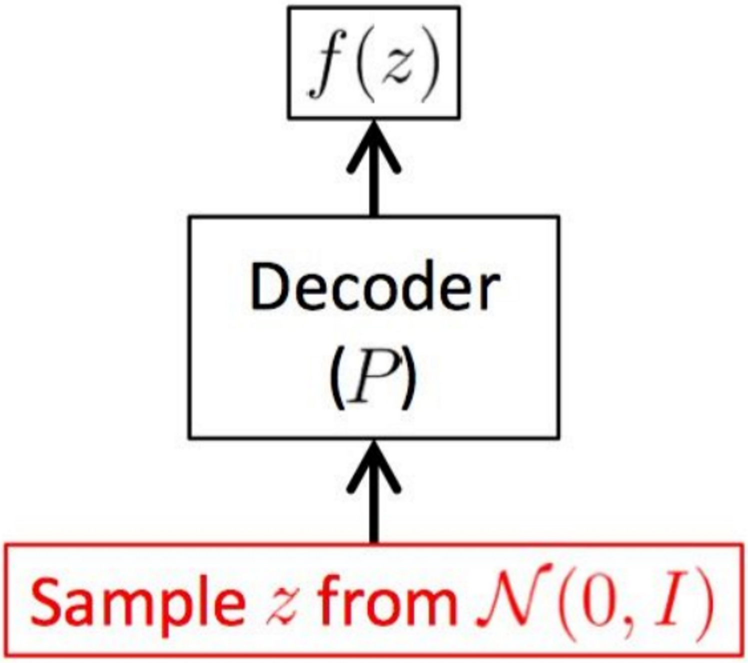
所以最后结果就是

$$KL\Big(N(\mu,\sigma^2)\Big\|N(0,1)\Big) = \frac{1}{2}\Big(-\log\sigma^2+\mu^2+\sigma^2-1\Big)$$

其中均值方差都是encoder网络的输出，$\mu=en_1(x)$, $\sigma=en_2(x)$

至此，就完成了VAE的整个loss计算过程，之后用梯度下降训练就可以了

训练完成后，VAE的生成过程如下图所示，即前面描述的z~p(z), x~p(x|z)生成过程 (z的先验预设为标准高斯分布)



g. 进一步理解

i. 为什么选择高斯分布？

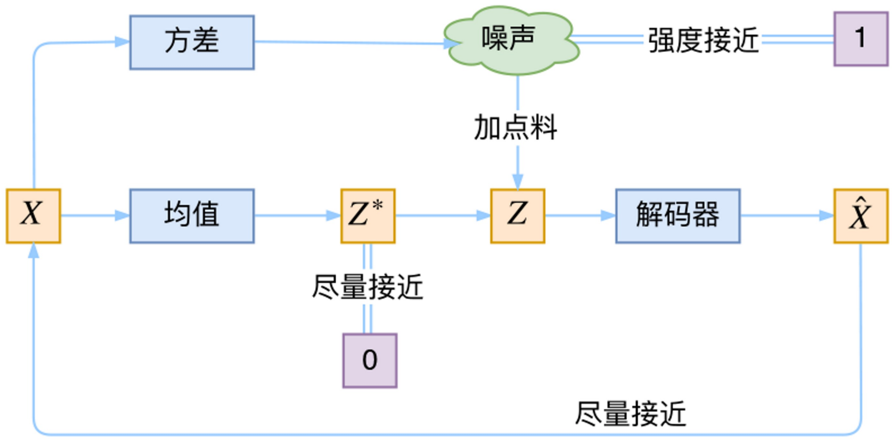
从KL计算的角度来分析，因为KL[p||q] = plogp/q, 如果在某个区域中概率为0的话，就会出现无穷大，使得数值计算不稳定，比如如果选择均匀分布，那么范围以外的点概率就是0，就会导致这个问题，所以应该选择一个在所有点概率都是非负的分布，因此高斯就是一个好的选择。

ii. encoder与decoder的相互牵制关系

encoder是要生成高斯分布的均值和方差的，虽然在上文分析中，这两个值的loss是从KL计算出来的，但是我们仍然想知道，这两个loss，会怎么影响模型的训练？随着训练模型的优化，loss会怎么改变？和decoder的loss又是什么样的关系？

首先要知道均值和方差的含义，这里要求高斯分布接近标准高斯分布，所以给定x后，z的均值应该接近0，方差应该接近1。均值为0，就是希望生成的隐变量与AE中的隐变量基本一致(en1(x)=μ(x)，encoder的这个输出起到了AE中encoder的作用)，就一个唯一的精确地包含可以重构样本的code，方差是用来添加噪声的，如果方差为0，那么VAE就变成了普通的AE，如果方差不为0，就是在原本的隐变量的基础上添加噪声，使得模型对噪声有鲁棒性，从而具有生成能力。所以KL项其实是相当于对encoder的一个正则项，希望它计算出来的z既有0均值，离原AE的编码不远，又有一定的鲁棒性。

下图就说明了方差和均值的变化是怎么影响整个模型的训练的



当decoder还没有训练好时（重构误差远大于KL loss），就会适当降低噪声（KL loss增加），使得拟合起来容易一些（重构误差开始下降）；反之，如果decoder训练得还不错时（重构误差小于KL loss），这时候噪声就会增加（KL loss减少），使得拟合更加困难了（重构误差又开始增加），这时候decoder就要想办法提高它的生成能力了。

因此，其实VAE中的encoder和decoder和GAN中的G和D异曲同工，也是有一定的对抗关系。

iii. 从联合分布的角度来推导VAE

前面VAE的推导是最小化后验分布的距离，可以尝试从联合分布距离来推导，即最小化KL[p(x, z) || q(x, z)]，使用一样的贝叶斯公式变换，到最后会发现推导出来的式子与前面一样，因此从联合分布的角度来理解和推导VAE也是可以的。推导过程详见https://kexue.fm/archives/5343

iv. 实践中的问题：其实decoder的p(x|z)也是分布

上一节在分析到"VAE的具体过程和loss"中，在分析loss函数形式时q(z|x)和q(z)就是分布，并且loss用的也是量度两个分布的KL距离，看到这我们就有疑问了，那为什么到了p(x|z)就不说分布了呢? 为什么直接说decoder直接输出x呢？下面就来解释一下生成模型p(x|z)的近似，到最后会发现其实正好是重构误差，前面直接说重构误差是为了直观地从VAE的作用来理解。

在近似q(z|x)和q(z)时，我们都选择了高斯分布，那么p(x|z)应该选择什么呢？在VAE原论文中给出了两种方案，高斯分布或者伯努利分布。

1) 伯努利分布

伯努利分布是一个二元模型，它只有两种取值，所以只适用于二元数值(比如MNIST数据集)

$$p(\xi) = \begin{cases} \rho, \ \xi = 1; \\ 1 - \rho, \ \xi = 0 \end{cases}$$

类似于encoder，这时候decoder应该计算伯努利分布的参数ρ，当x有D维时，可以算得

$$q(x|z) = \prod_{k=1}^{D} \left(\rho_{(k)}(z)\right)^{x_{(k)}} \left(1 - \rho_{(k)}(z)\right)^{1-x_{(k)}}$$

两边取log，就得到了

$$-\ln q(x|z)=\sum_{k=1}^{D}\left[-x_{(k)}\ln\rho_{(k)}(z)-(1-x_{(k)})\ln\left(1-\rho_{(k)}(z)\right)\right]$$

看，这是不是就是我们熟悉的交叉熵loss，因此在具体实现上，当x的取值是二值时，需要在decoder最后加一层使得输出在[0,1]之间，与x的取值范围一致，比如可以加一层sigmoid，然后使用交叉熵作为loss。

2) 高斯分布

一般情况下，x的取值不会是二元取值这么简单，一般来说，大部分事物的分布规律都可以用高斯分布来模拟。

高斯分布的函数

$$q(x|z)=\frac{1}{\prod_{k=1}^{D}\sqrt{2\pi\tilde{\sigma}_{(k)}^{2}(z)}}\exp\left(-\frac{1}{2}\left\|\frac{x-\tilde{\mu}(z)}{\tilde{\sigma}(z)}\right\|^{2}\right)$$

这时候decoder输出的就应该是均值μ(z)和方差σ(z)。两边取log，就得到了

$$-\ln q(x|z)=\frac{1}{2}\left\|\frac{x-\tilde{\mu}(z)}{\tilde{\sigma}(z)}\right\|^{2}+\frac{D}{2}\ln 2\pi+\frac{1}{2}\sum_{k=1}^{D}\ln\tilde{\sigma}_{(k)}^{2}(z)$$

一般σ会固定为常数，因此这个loss就只和均值μ(z)有关，就变成了

$$-\ln q(x|z)\sim\frac{1}{2\tilde{\sigma}^{2}}\left\|x-\tilde{\mu}(z)\right\|^{2}$$

在encoder中，均值μ(x)希望与普通AE的编码z一致，于是，同样地，因为VAE同样是用于重构的，所以μ(z)希望与原输入x一致(在VAE中，虽然说的都是分布，但是其实都是很窄的高斯分布，所以生成的差不了多少)，所以上面的loss其实就是MSE，重构loss

3) 一个常见的误解

既然p(x, z) = p(x|z)p(z)，p(z)是服从标准高斯分布的先验，现在又说p(x|z)是高斯分布，那两个高斯相乘，不就得到了样本分布p(x, z)也是高斯分布吗？可是它明明是个复杂分布呀

- 来看一下这个想法的错误在哪里。

  p(z)是关于**z**的标准高斯分布，也即

  $$\mathrm{p(z)}=\frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}$$

  p(x|z)是关于**x**的高斯分布，它的均值μ(z)和方差σ(z)是由z通过decoder计算得到的，也就是z的函数

  $$\mathrm{p(x|z)}=\frac{1}{\sigma(\mathrm{z})\sqrt{2\pi}}e^{-\frac{(x-\mu(\mathrm{z}))^2}{2\sigma(z)^2}}$$

  那这两个分布相乘是高斯吗？不是，如果是关于(x,z)的高斯分布，那么它必须能够写成x的二次型和z的二次型，而$\mu(\mathrm{z})$和$\sigma(\mathrm{z})$是通过神经网络计算得到的，是z的一个复杂函数，它没法写成z的二次型，所以这是一个误解，这两个高斯分布相乘并不是关于x和z的高斯分布，就是一个复杂的分布。

v. 关于采样

疑问：既然$q(z_i|x_i)$是一个分布，那么对于每个数据点x_i，我是不是要采样好多个z才能保证学习得到分布呢?
答案：不是，一个就够了。也正是因为这个原因，所以VAE的实现看起来与AE相差不大。为什么只需要采样一个呢？可以这么理解，训练会运行很多个epoch，每一次都采样一个，相当于采样了好几次嘛，所以并不需要每次采样多个来保证覆盖分布，而实验也证明了采样一个或者多个结果没有差异。第二个原因是，$q(z|x)$是一个方差比较小的分布，很窄的高斯分布，所以每次采样出来的值相差不大。（思考：所以比起GAN，它的多样性比较弱，不会生成什么比较amazing的东西？）