Variational AutoEncoder

变分自动编码器（VAE）

汇报人脑瘫警告！

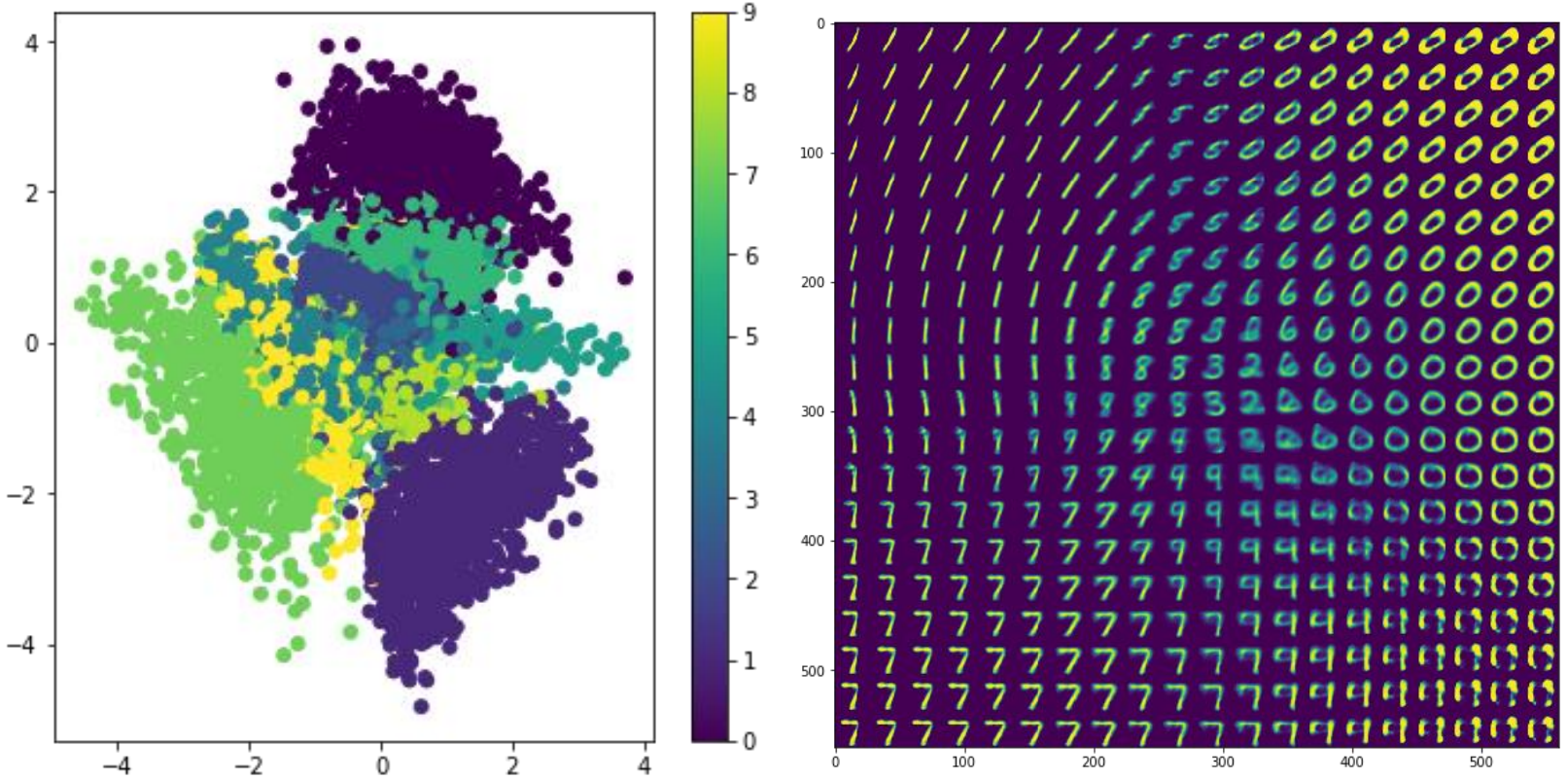
- 由于汇报人脑瘫严重，学VAE的时候是从概率论开始学的
- 因此后面的内容可能有很多大家已经学过的、但是又被拿出来一通讲
- 请原谅

目录

- 编码器、自动编码器、变分自动编码器原理
- VAE代码与相关trick
- 数学部分

声明

- VAE的本职工作是数据压缩和数据生成，我们将据此引入VAE。
- VAE的分类功能最后再讲。



符号

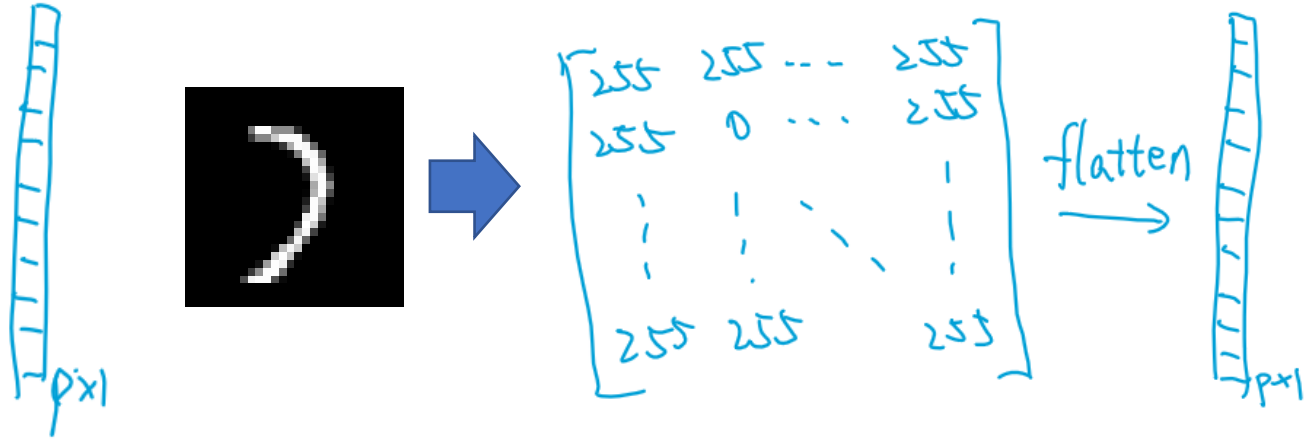
训练集：$\underline{X}$、$\underline{Y}$

$\underline{X} = (x^{(1)}, x^{(2)}, \cdots, x^{(n)})_{p \times n}$

$\underline{Y} = (y^{(1)}, y^{(2)}, \cdots, y^{(n)})_{1 \times n}$

[7, 9, 8, 6, 8]

$x^{(i)} =$



$y^{(i)} \in R$

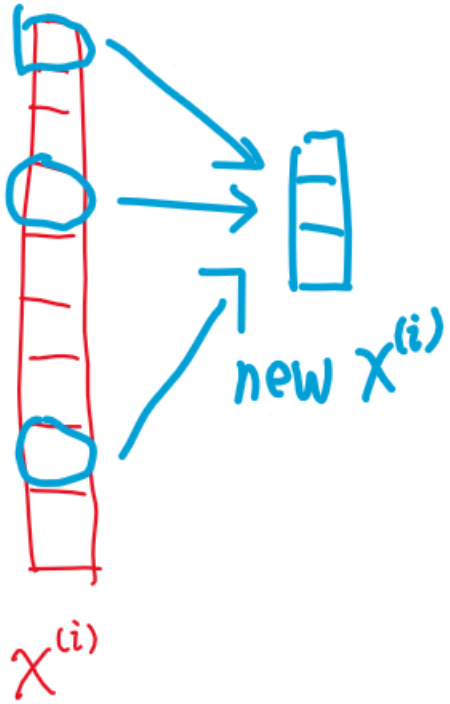
降维（数据压缩）

- selection
  - 从原来的高维特征向量中选出一些组成低维特征向量

- extraction
  - 原来的高维特征向量经过计算转换为低维特征向量

降维 – 脑瘫法

- 从原来的高维向量中选一些组成低维向量(selection)



降维 – 线代法

- PCA：主成分分析(extraction)

$$C_{p\times p} = \frac{1}{n-1} X_{p\times n} X^T_{n\times p}$$

$$C = P_{p\times p} \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_p \end{bmatrix} P^T_{p\times p}$$

其中λ_i按$\lambda_1 > \lambda_2 > \cdots > \lambda_p$排列

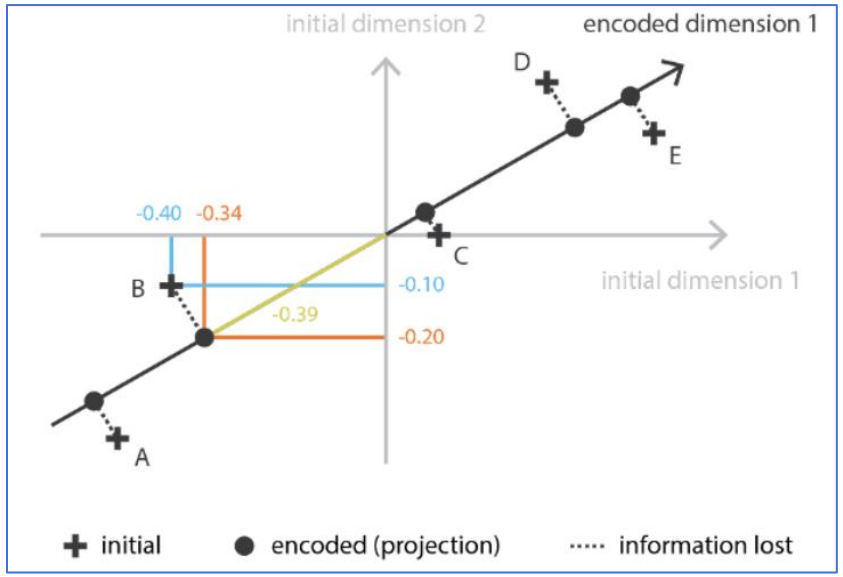
要想降至m维（$m < p$）

$$newX_{m\times n} = (P\text{的前}m\text{列})^T_{m\times p} \cdot X_{p\times n}$$

要想还原回p维

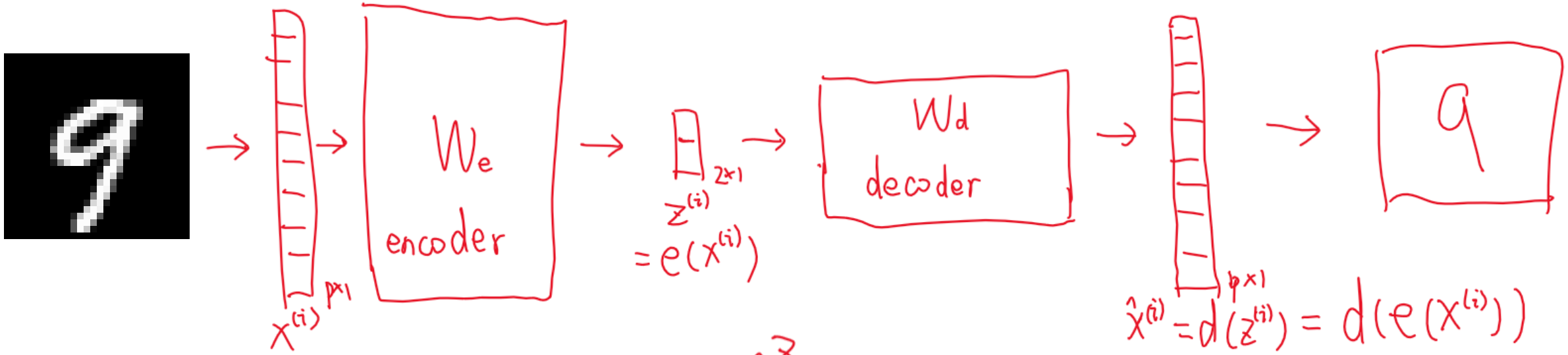
$$newnewX_{p\times n} = (P\text{的前}n\text{列})_{p\times m} \cdot newX_{m\times n}$$

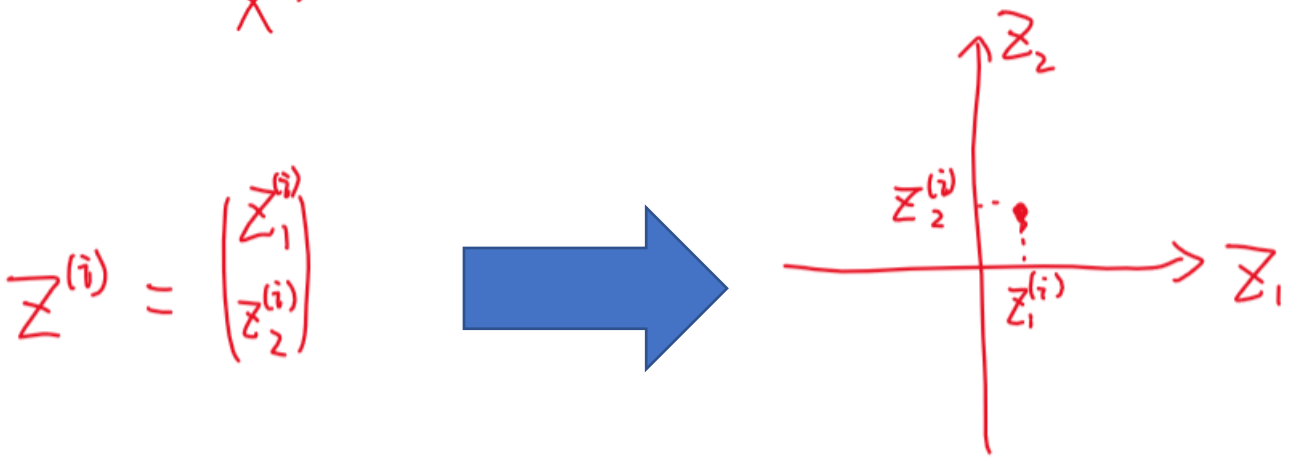
右图即为把二维数据降成一维：



分割线

以上都是手工方法，接下来全是神经网络方法
提示：接下来的例子全都是：图片->高维向量->二维向量（平面上的点）->高维向量->图片



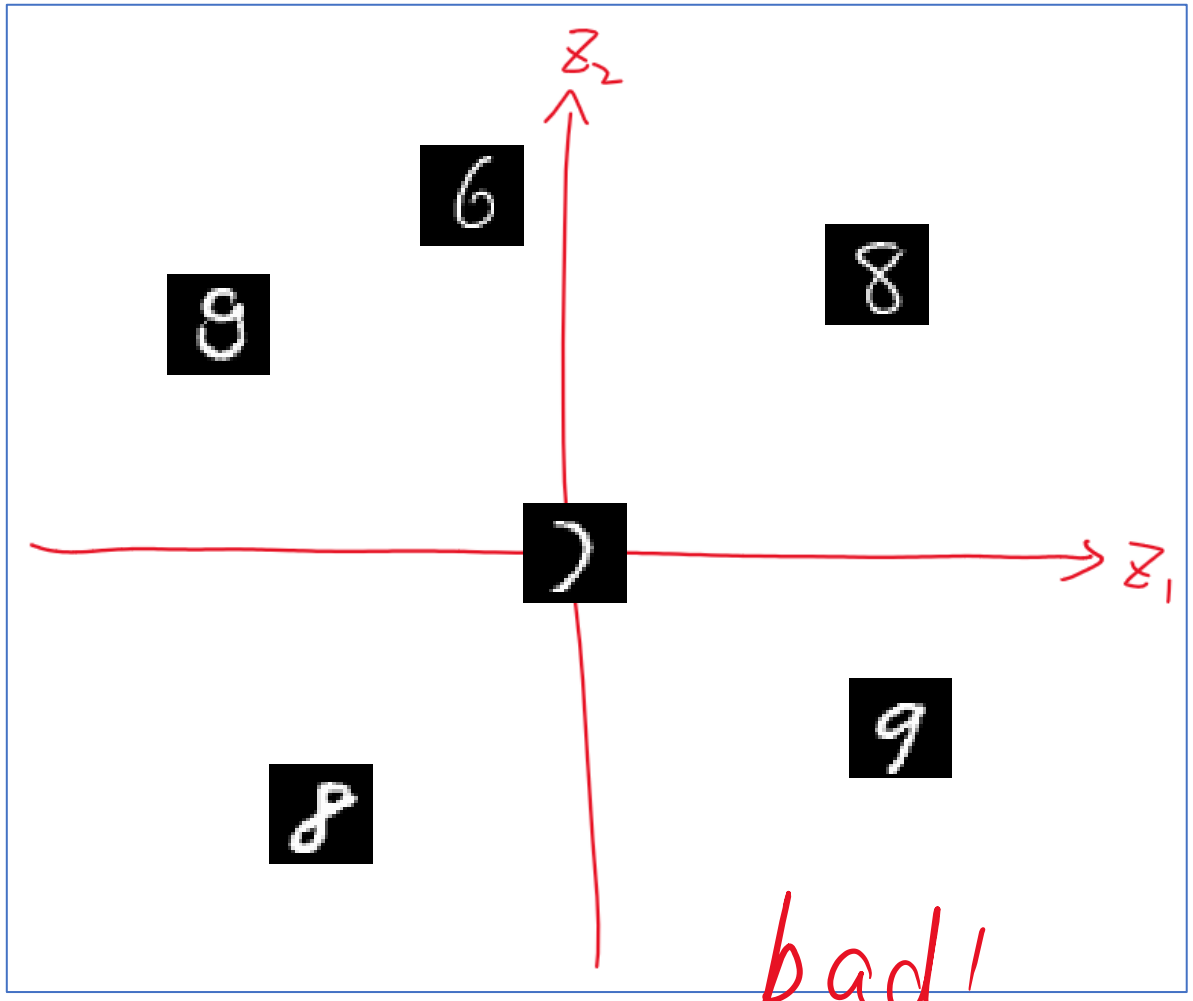


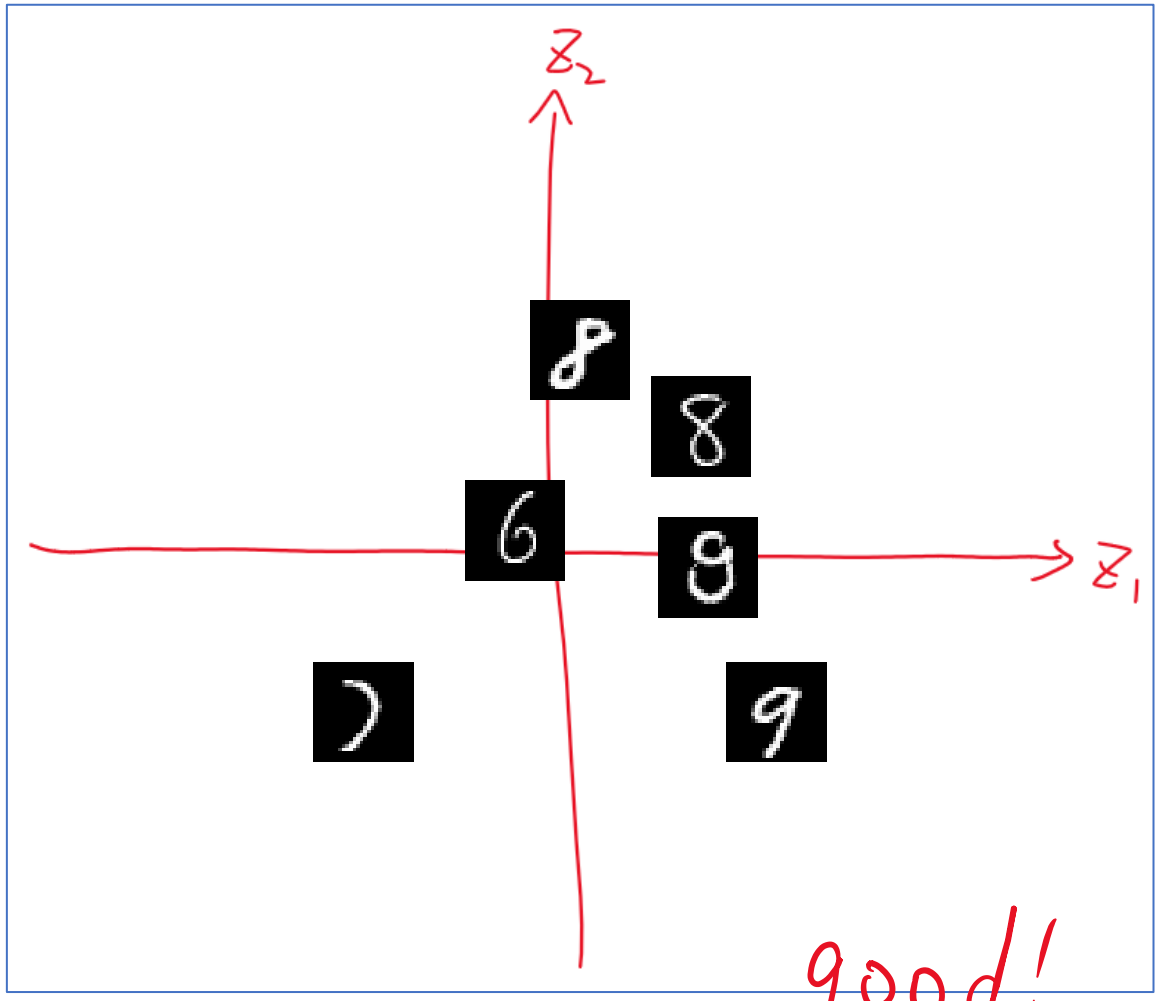
分割线

- 以上是手工方法，接下来是神经网络方法

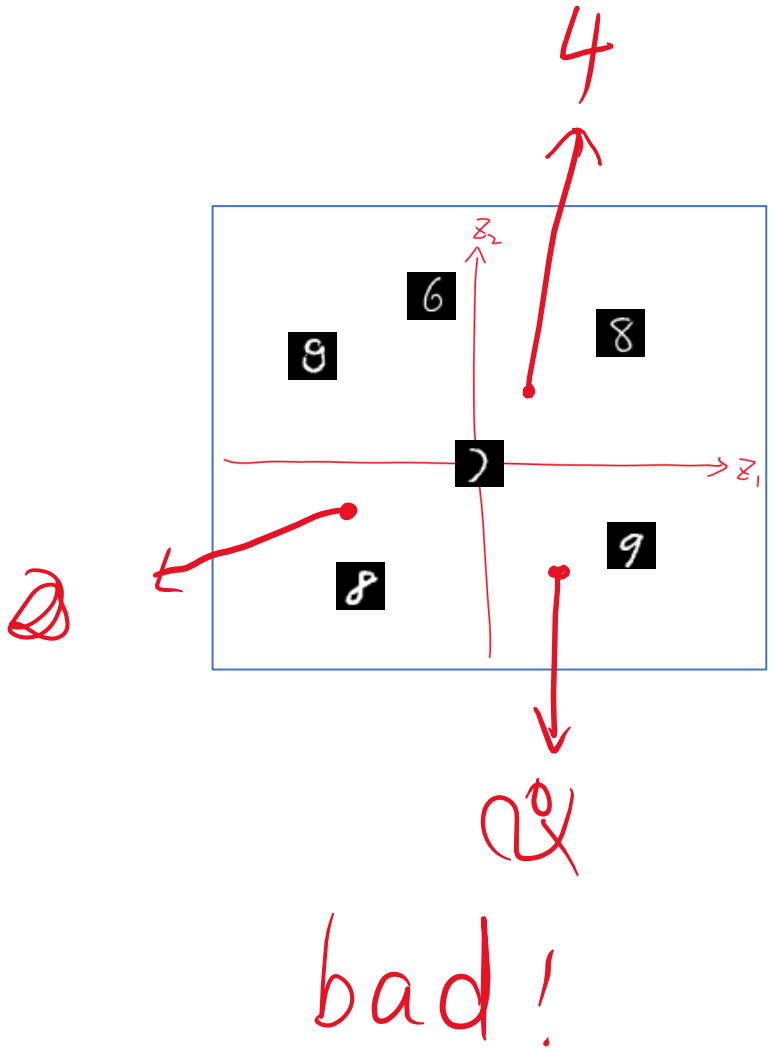
- 我们将从
  - 隐层表示的质量
  - 数据生成的质量
- 来评估模型

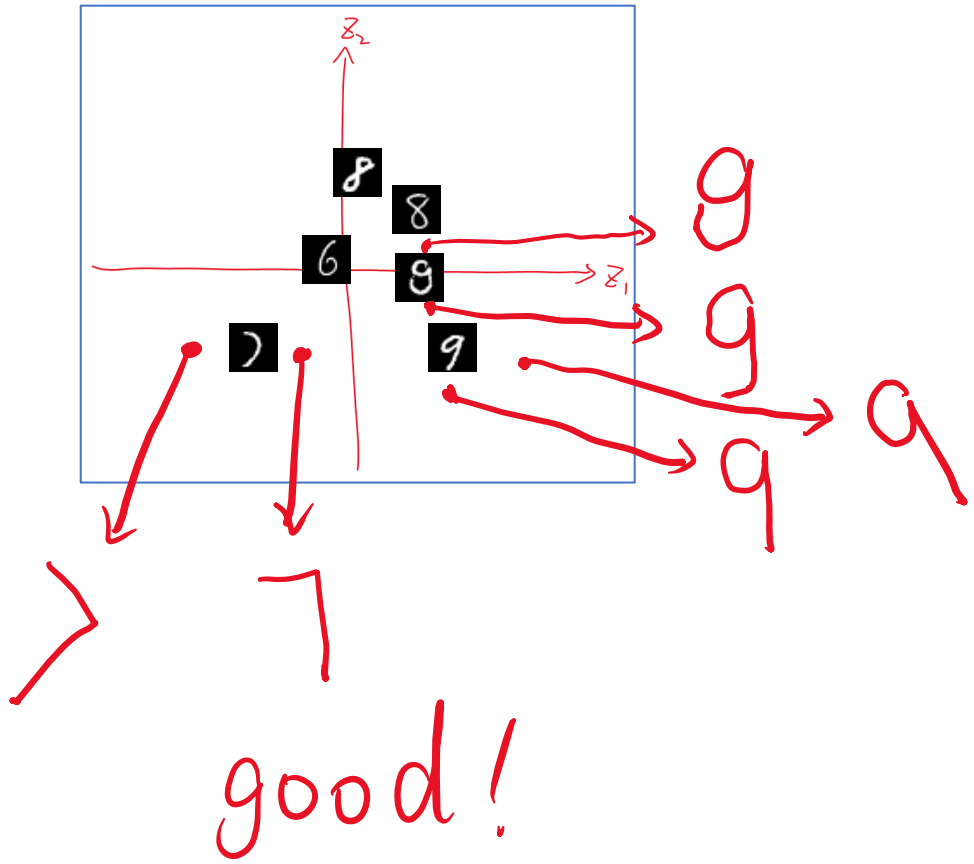
隐层表示的质量



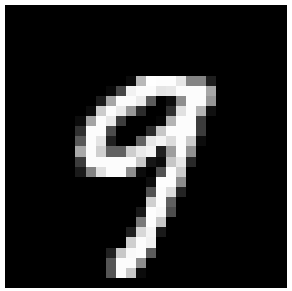


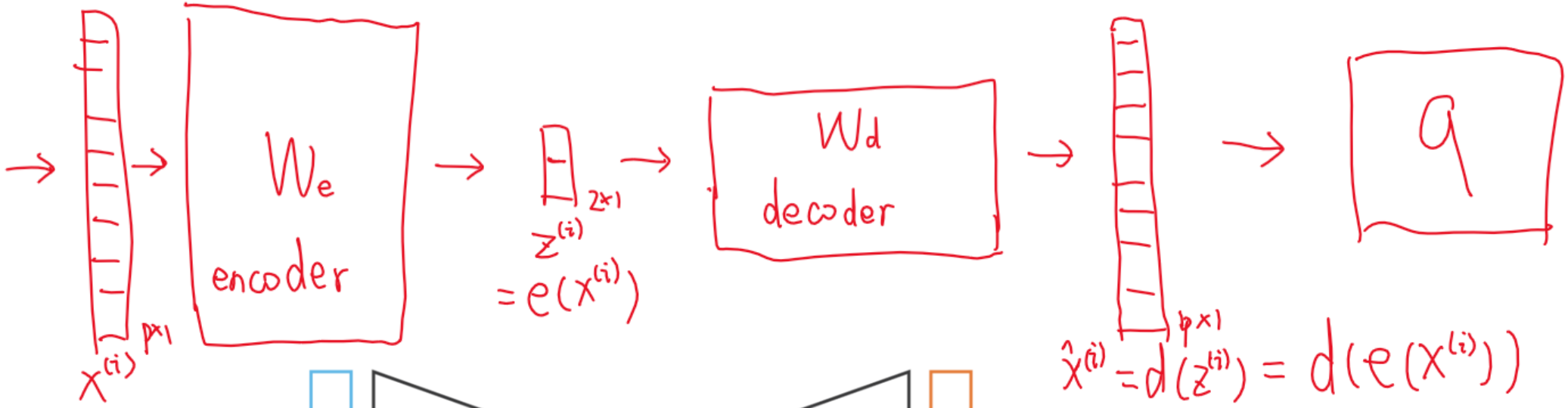
数据生成的质量

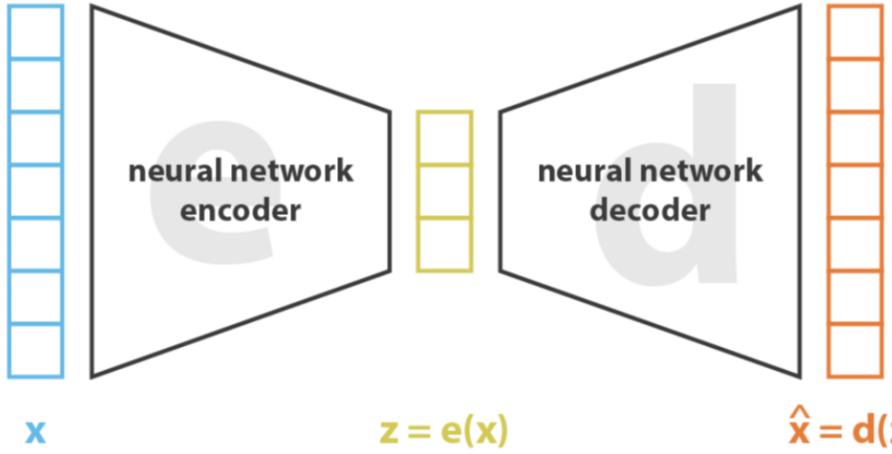




AutoEncoder – 模型







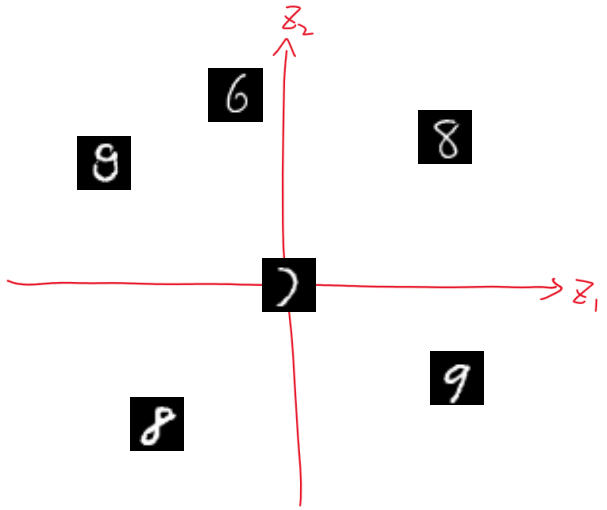
$$\text{loss} = \| x - \hat{x} \|^2 = \| x - d(z) \|^2 = \| x - d(e(x)) \|^2$$

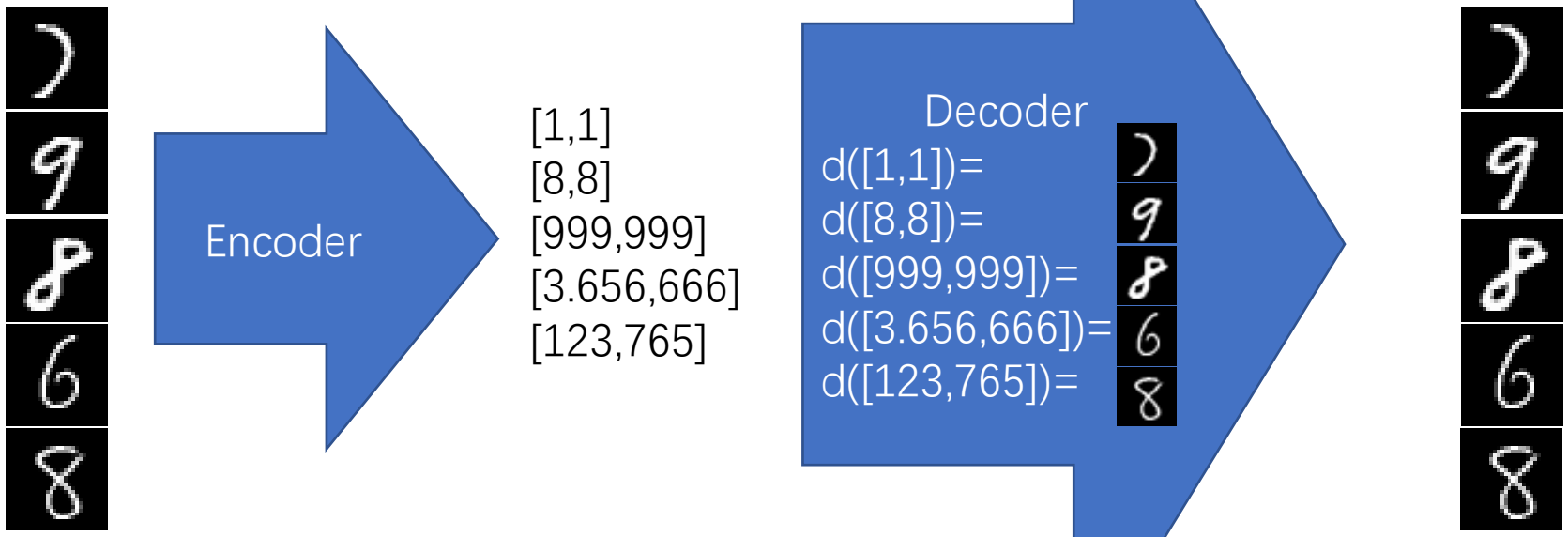
AutoEncoder – 隐层表示 - 自由自在

如果编码器和解码器够大，很容易造成过拟合，loss为0的那种

所有图片的像素信息都存在解码器里，编码器就是给输入图片找到它对应的编号，解码器就根据编号找到图片的像素信息，直接输出出来。可以想象：编码器编的编号有没有意义都无所谓、大小也无所谓，反正一切像素信息（最基础的、最不抽象的信息）都保存在解码器里了。

隐空间里的点如同天上的繁星，色泽鲜艳，相距甚远

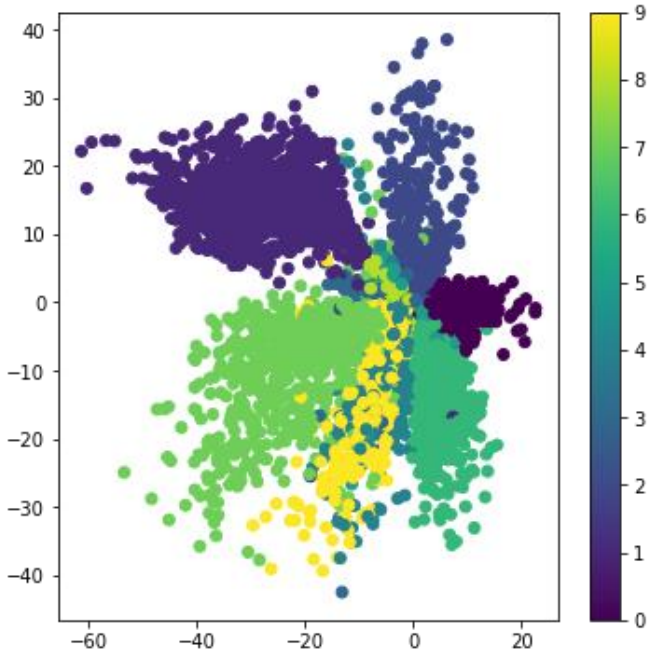


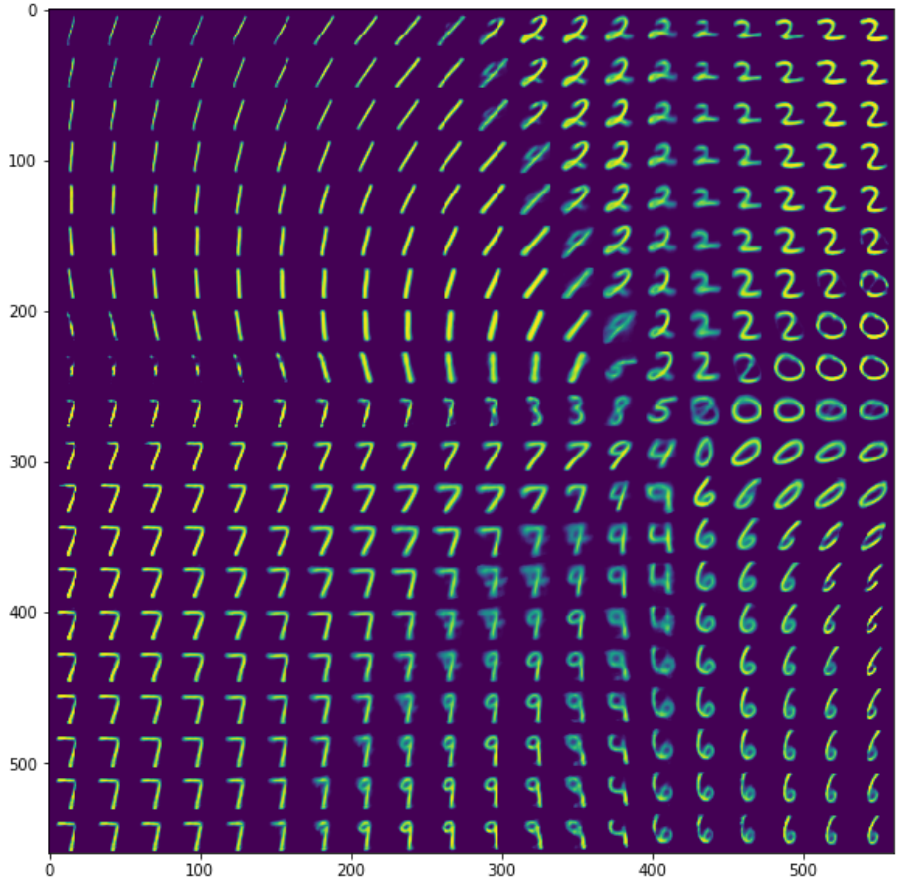


loss=0!

AE训练结果举例

由于模型不是非常的大，所以过拟合也不是非常严重，但还是能看出过拟合（比如下图中的大豁口）






```
Epoch 100/100
60000/60000 [==============================] - 23s 385us/step - loss: 125.0440 - val_loss: 133.9662
```

解决过拟合

- 限制编码空间的范围
- 减小encoder和decoder的体积

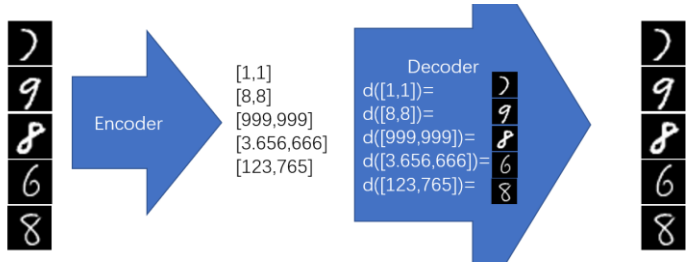
为什么要解决过拟合

希望提高模型生成数据的能力

如何评价模型过拟合的程度

观察生成数据的质量

解决过拟合-限制编码空间的范围



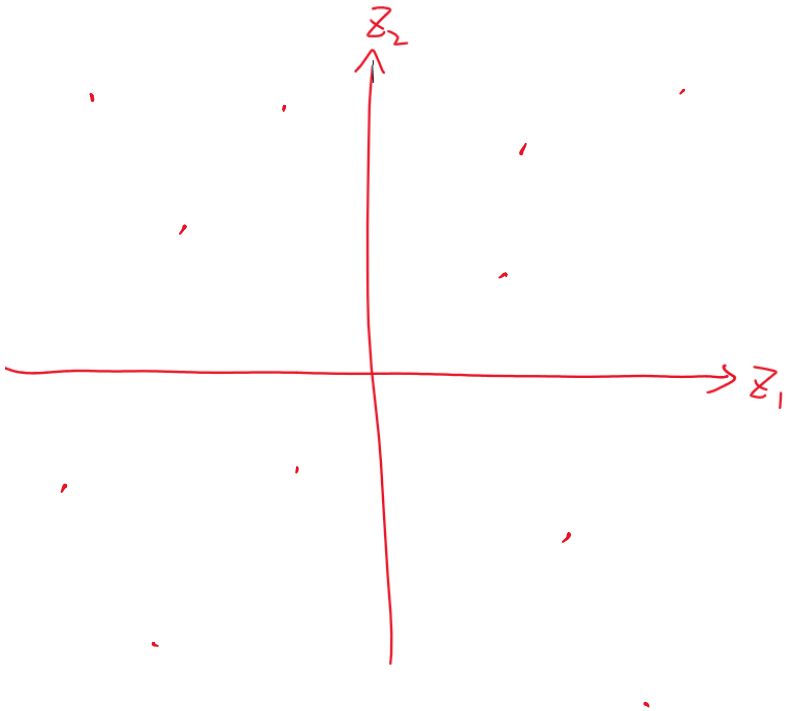
没有意义。

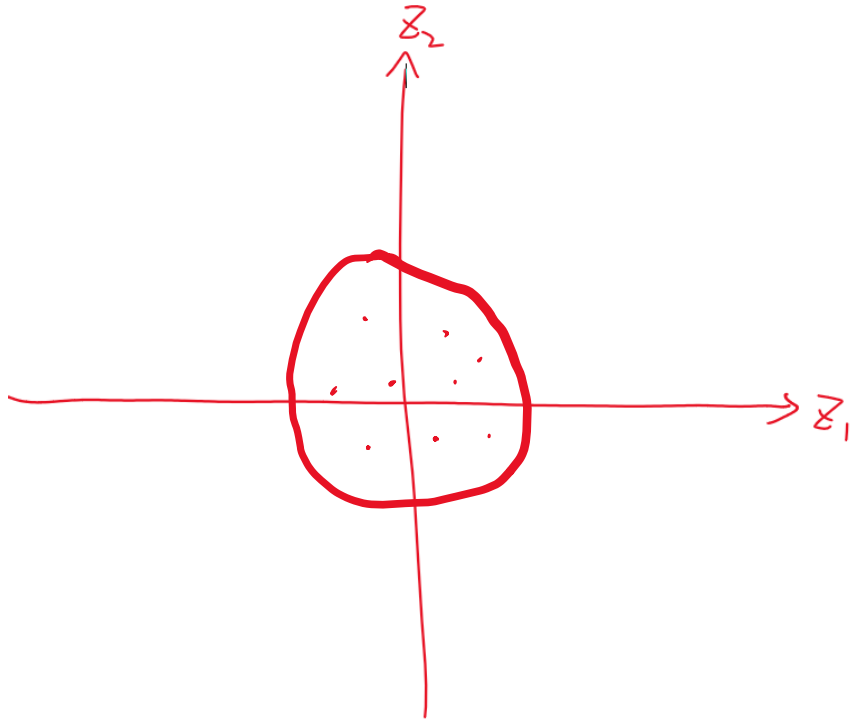
因为编解码器足够大的话，就能把每个样本任意的映射到点上，然后用巨大的

就比如：所有图片的像素信息都存在解码器里，编码器就是给输入图片找到它

找到图片的像素信息，直接输出出来。

可以想象：编码器编的编号有没有意义都无所谓、大小也无所谓，反正一切像素信息（最基础的、最不抽象的信息）都保存在解码器里了。



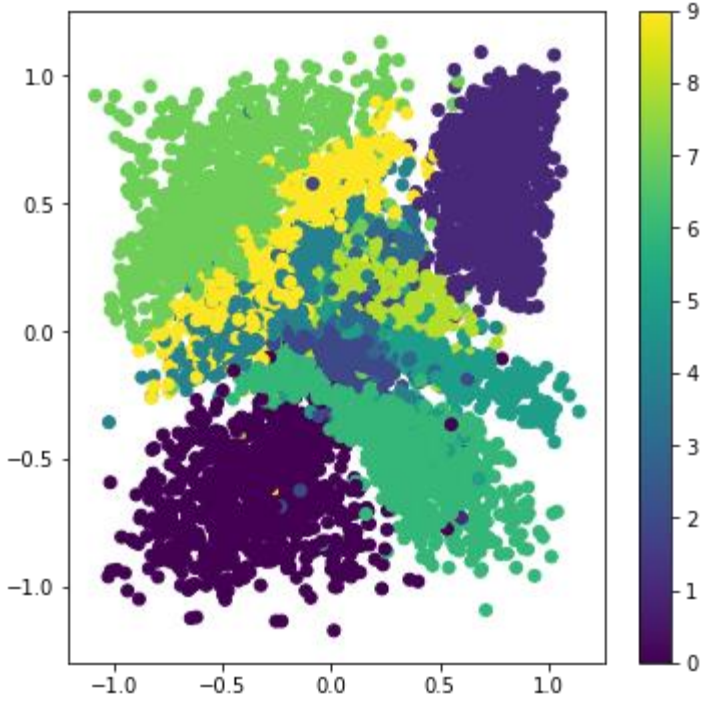


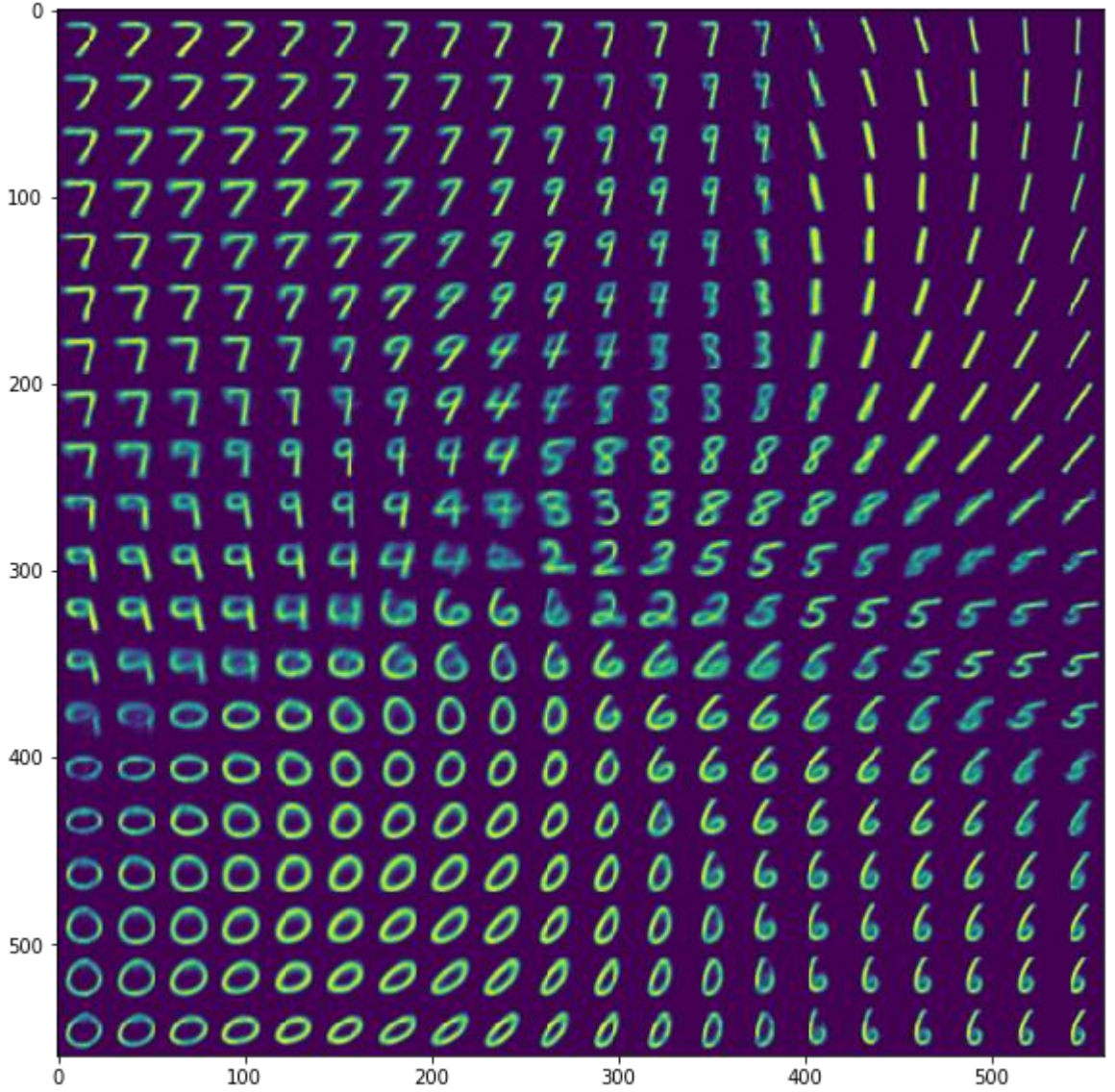
而且，无论是多小的平面，都可以容纳无数个谁也不搭理谁的点，只是个放缩的问题罢了。。

注意！再次提示

由于实验模型不是非常的大，所以过拟合也不是非常严重（毕竟实验结果里loss也没降到0，甚至都不能说是很低），但从图像上还是能看出过拟合，请大家一定要想办法看出过拟合！（在减轻过拟合的实验中甚至直接欠拟合了。。）

解决过拟合-限制编码空间的范围

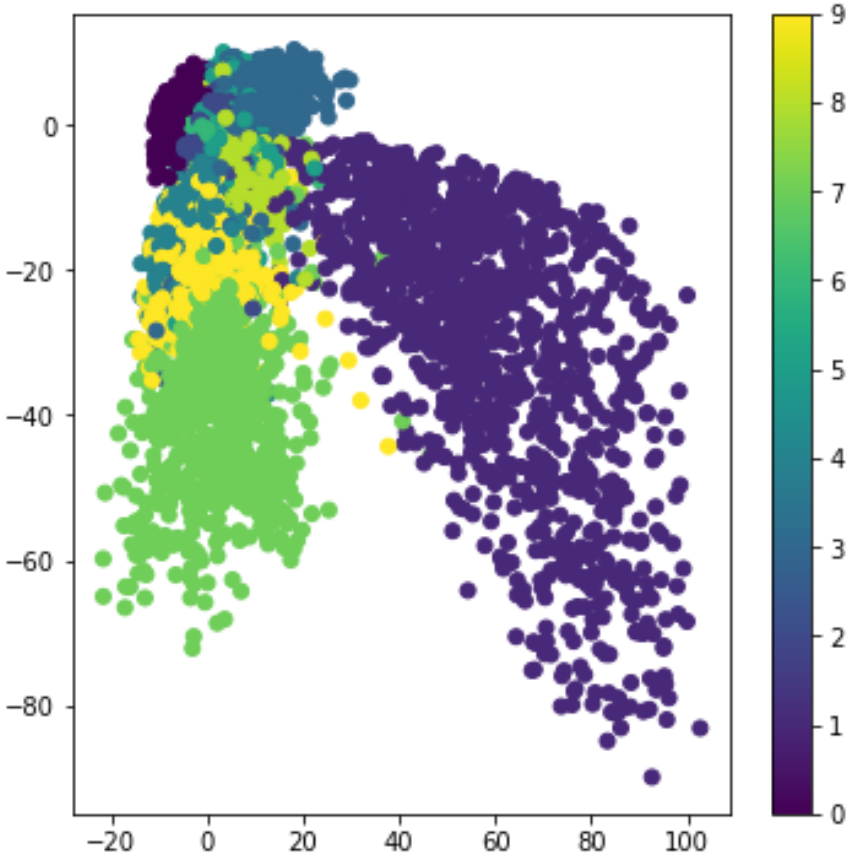


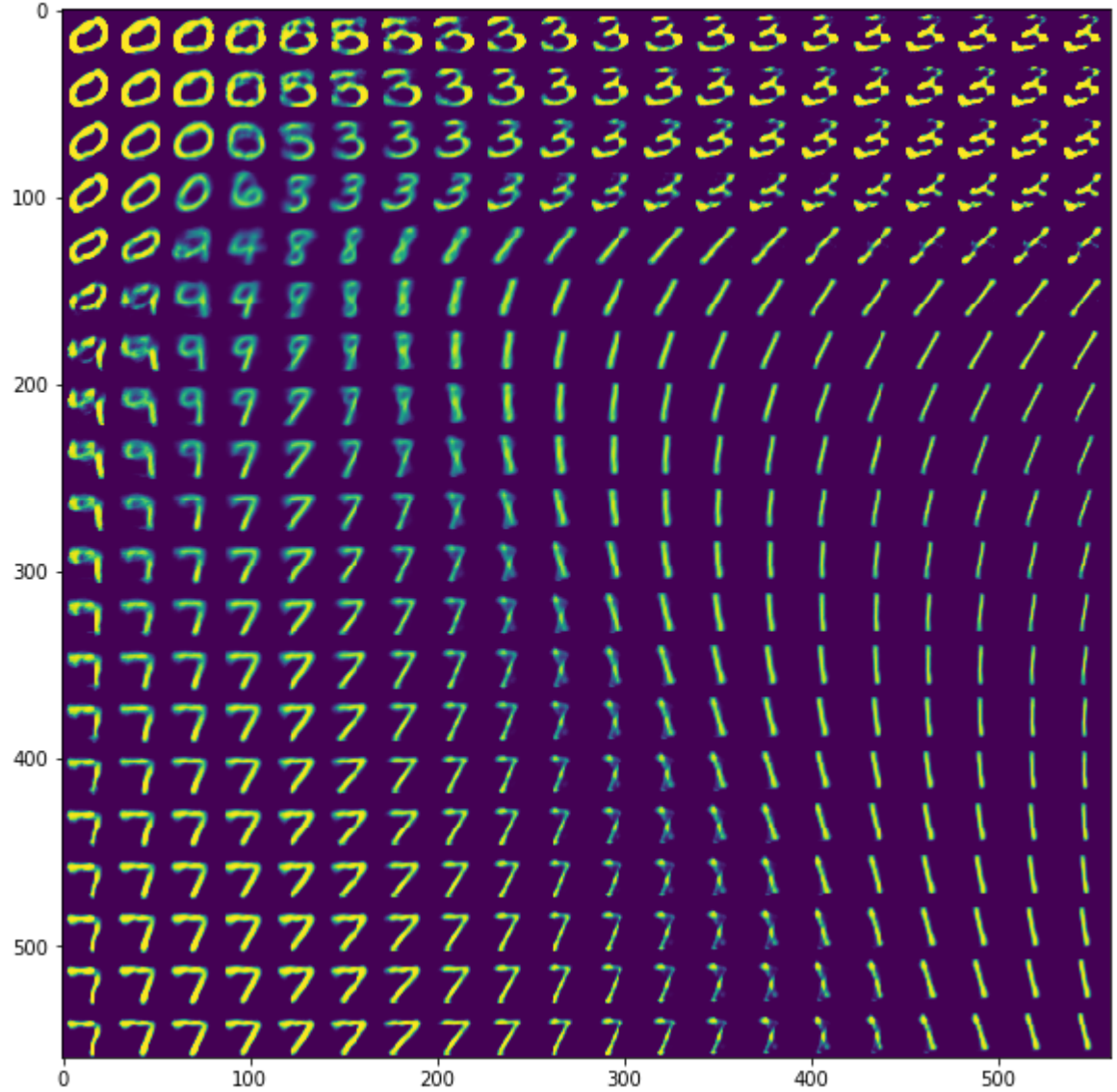



```
Epoch 100/100
60000/60000 [==============================] - 26s 439us/step - loss: 129.7683 - val_loss: 133.9667
```

解决过拟合-减小Encoder和Decoder的体积

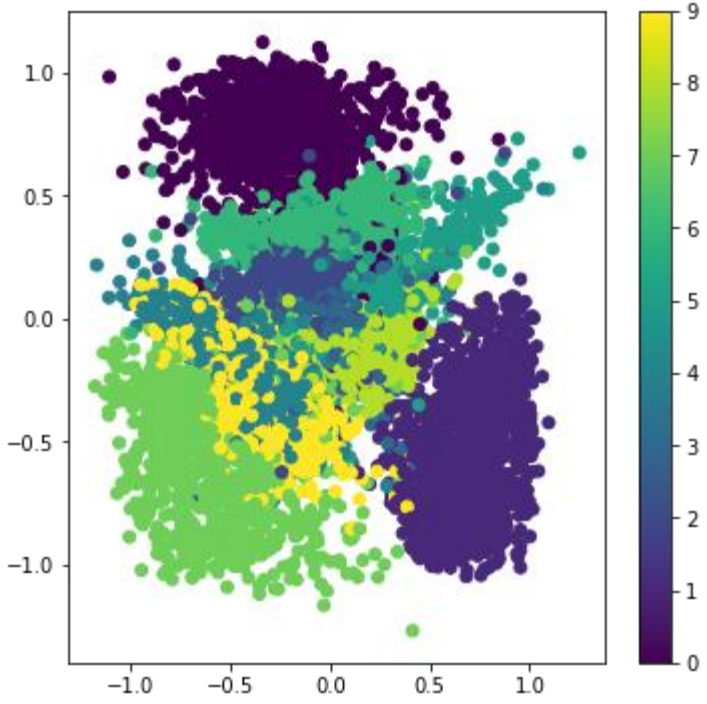
本来模型就不大，现在都有点欠拟合了

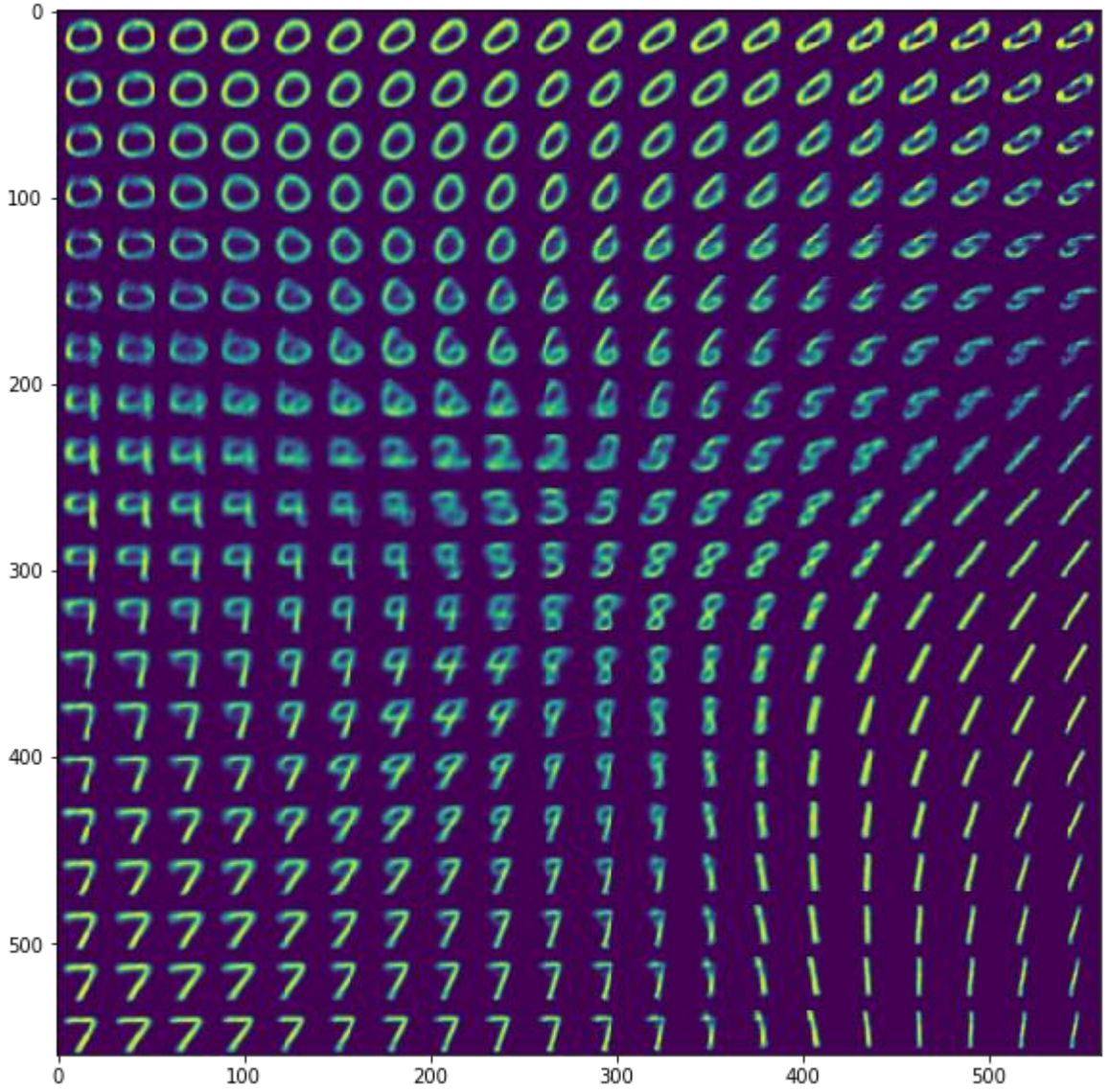





```
Epoch 50/50
60000/60000 [==============================] - 11s 187us/step - loss: 140.6021 - val_loss: 144.9442
```

限制范围+减小EncoderDecoder体积

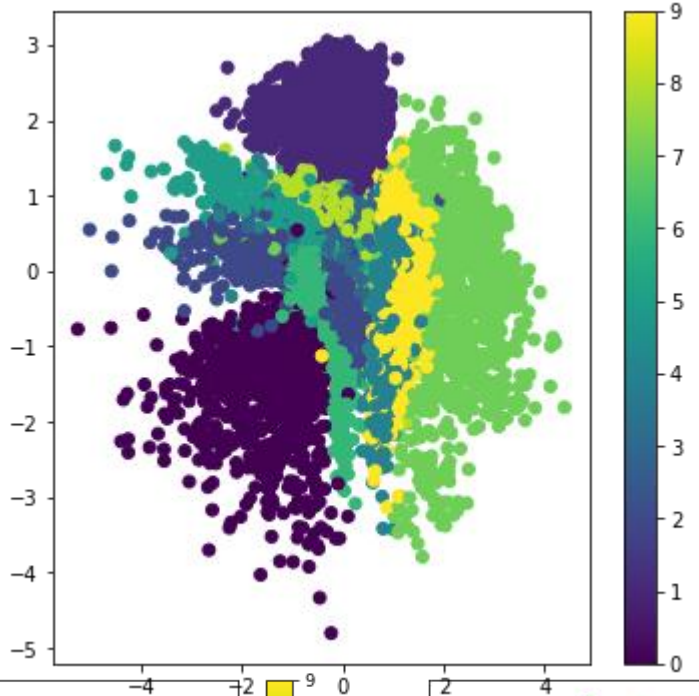




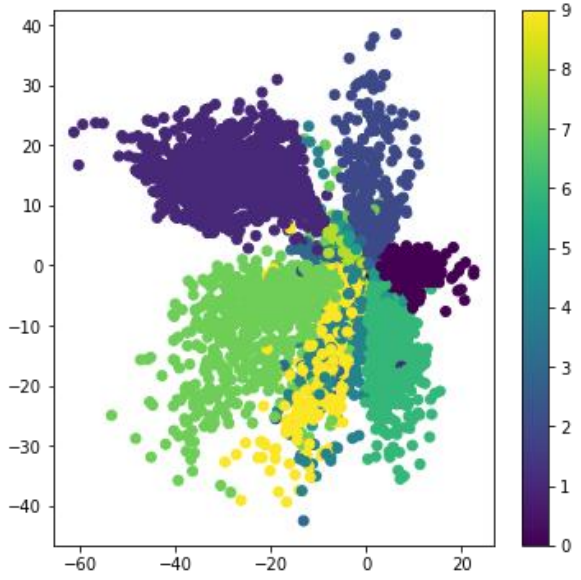

```
Epoch 50/50
60000/60000 [==============================] - 5s 90us/step - loss: 141.7742 - val_loss: 145.4607
```

实验结果总结

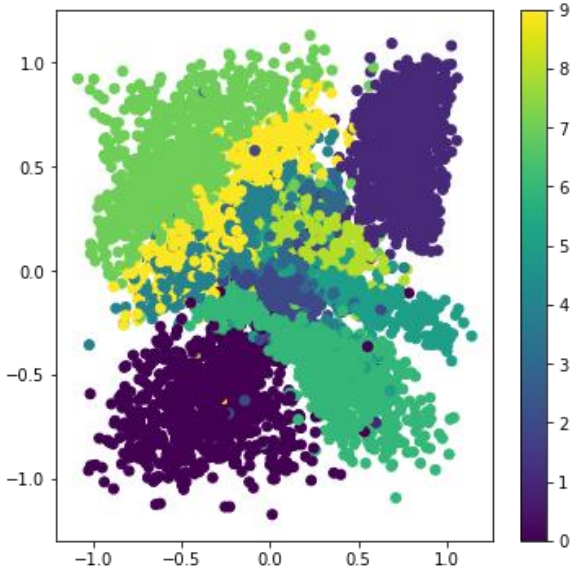
模型	训练轮数	train loss	validate loss
AE	100	125.0440	133.9662
AE-constrained	100	129.7683	133.9667
AE-small	50	140.6021	144.9442
AE-small-constrained	50	141.7742	145.4607



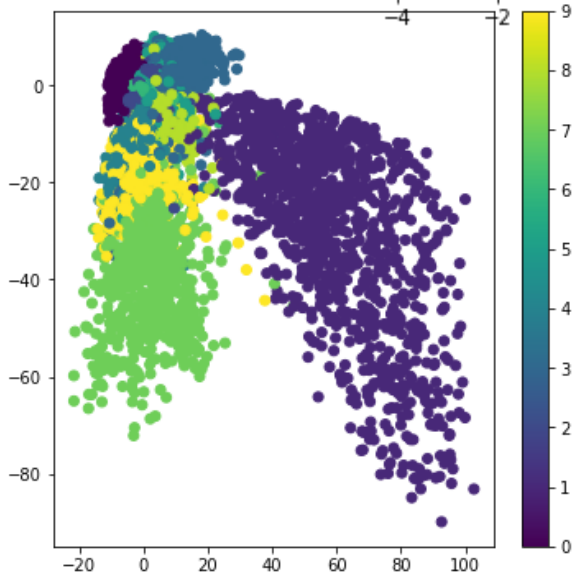
VAE



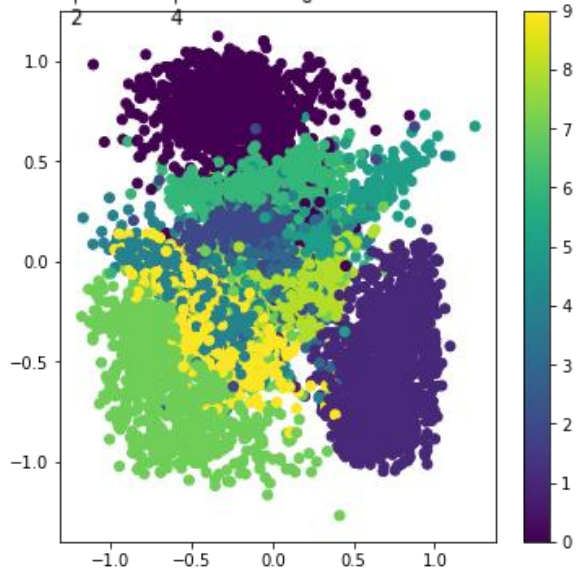
AE



AE-constrained



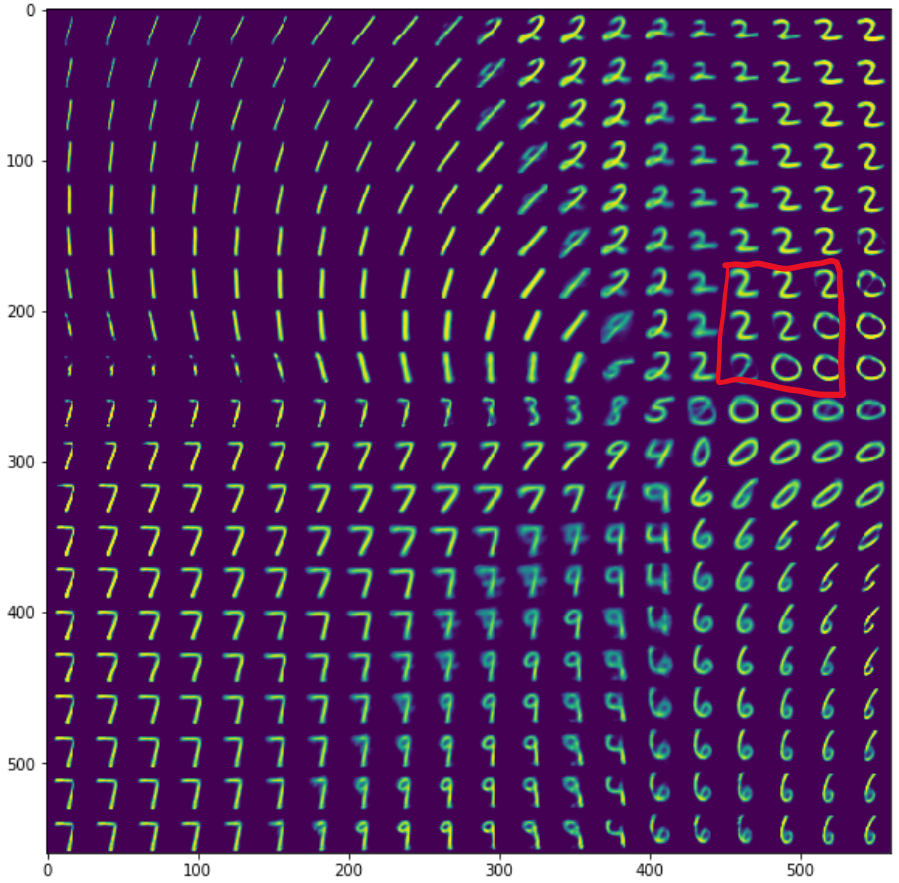
AE-small



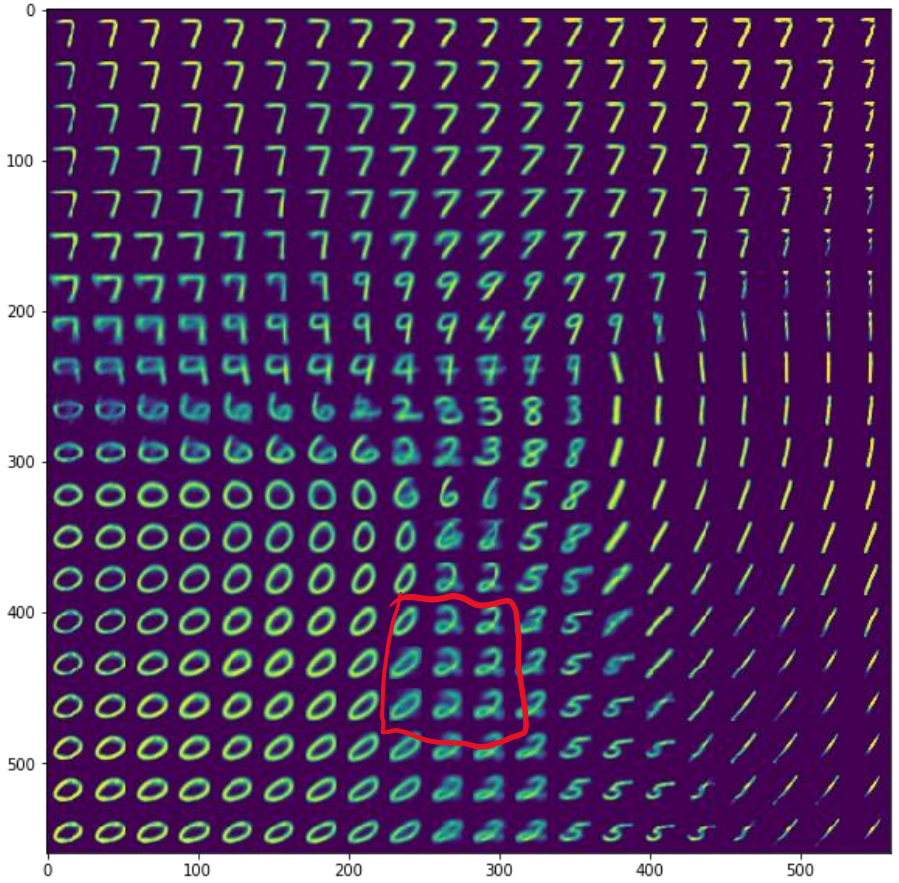
AE-small-constrained

AutoEncoder隐空间可解释性 – 差

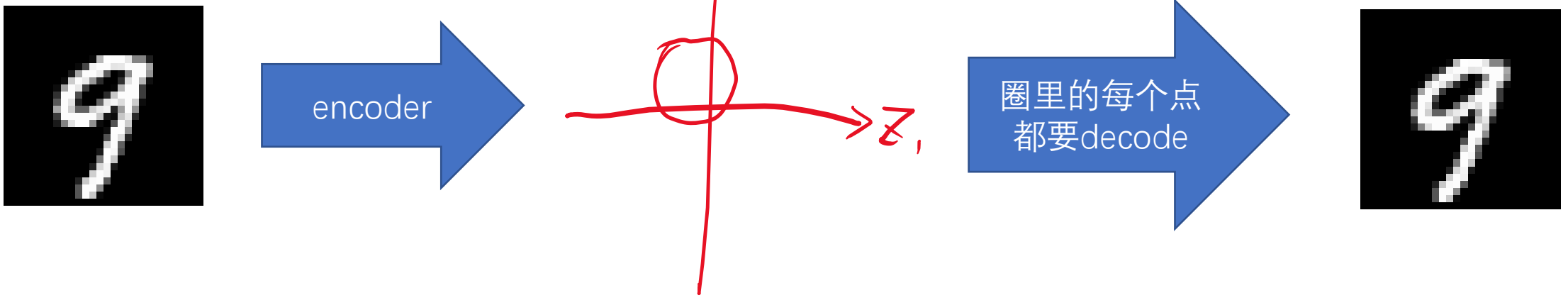
由于我的实验中大模型也比较小，过拟合不是非常严重，但也存在生成的图片随z移动而突变的问题



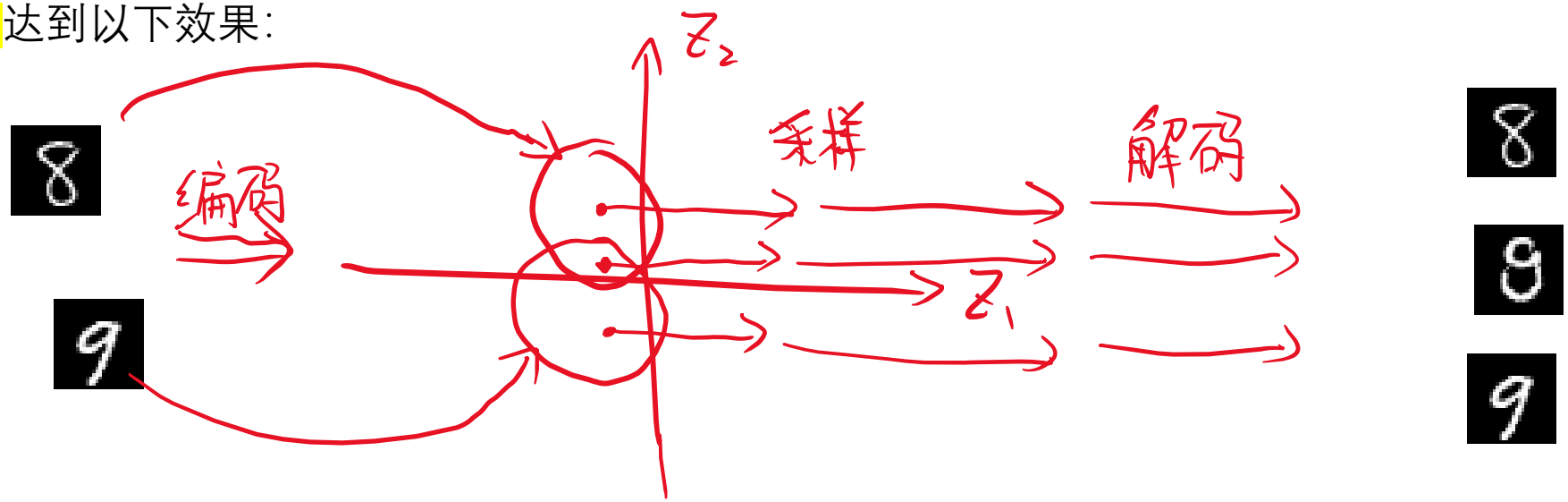
理想中的隐空间 – 一个轴对应数字、一个轴对应旋转、一个轴对应大小…太美好了



VAE：将每张图片编码为一个分布



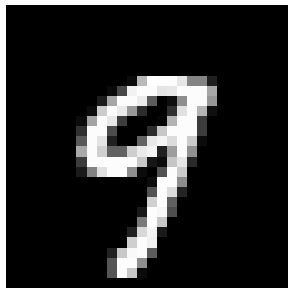
我们构造的损失函数要求：越靠近圈中间，重构后的图片就必须越像原图片，四周的差点就差点，但是也得像。
这样就容易达到以下效果：



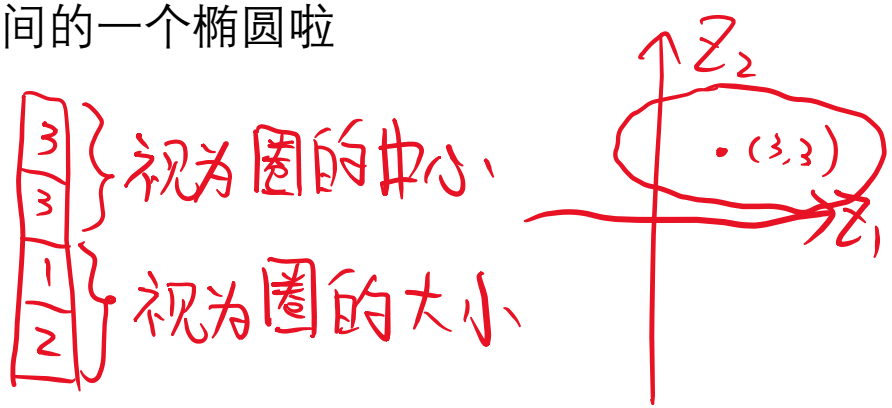
这种损失函数迫使：要想loss低，两张图片对应的圈的重叠部分采样的z解码后必须和两个图片都像

如何让一个图片对应隐空间的一个面？
如何让圈里的每个点（无数个点）都解码算加权的loss？

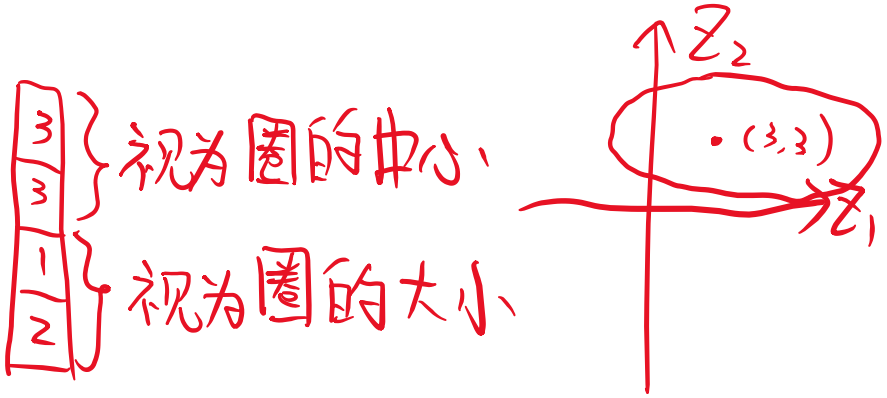
第一个问题的答案：
让一个图片encode成一个四维向量：分别表示二维高斯分布的均值和方差（假设二维高斯的二维独立）
这样一个图就可以算作编码为隐空间的一个椭圆啦

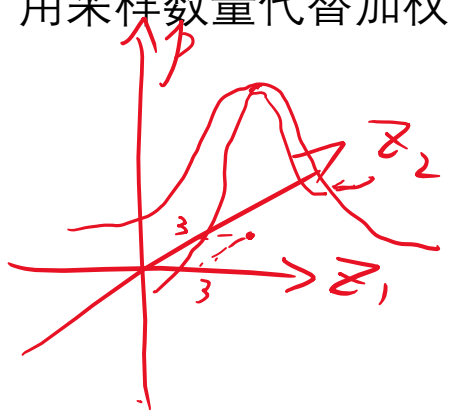


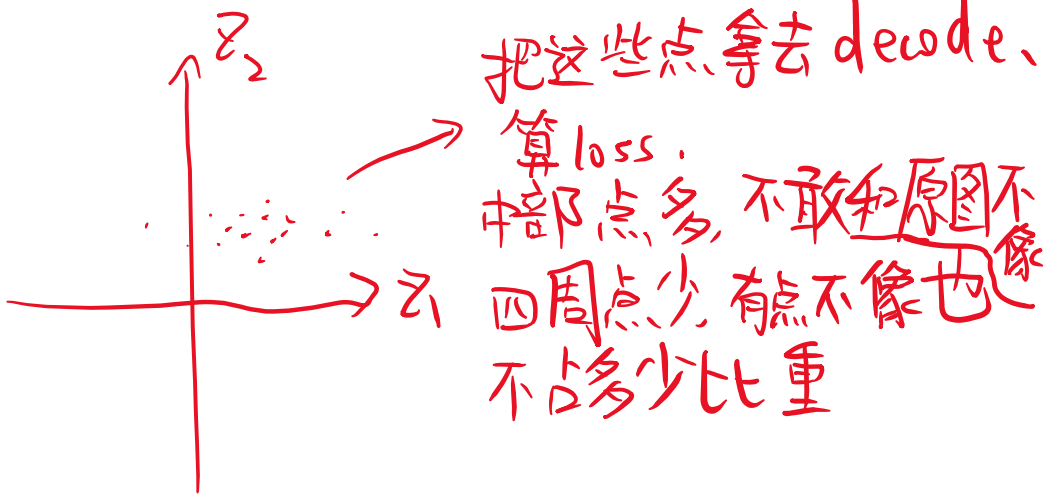
Encoder

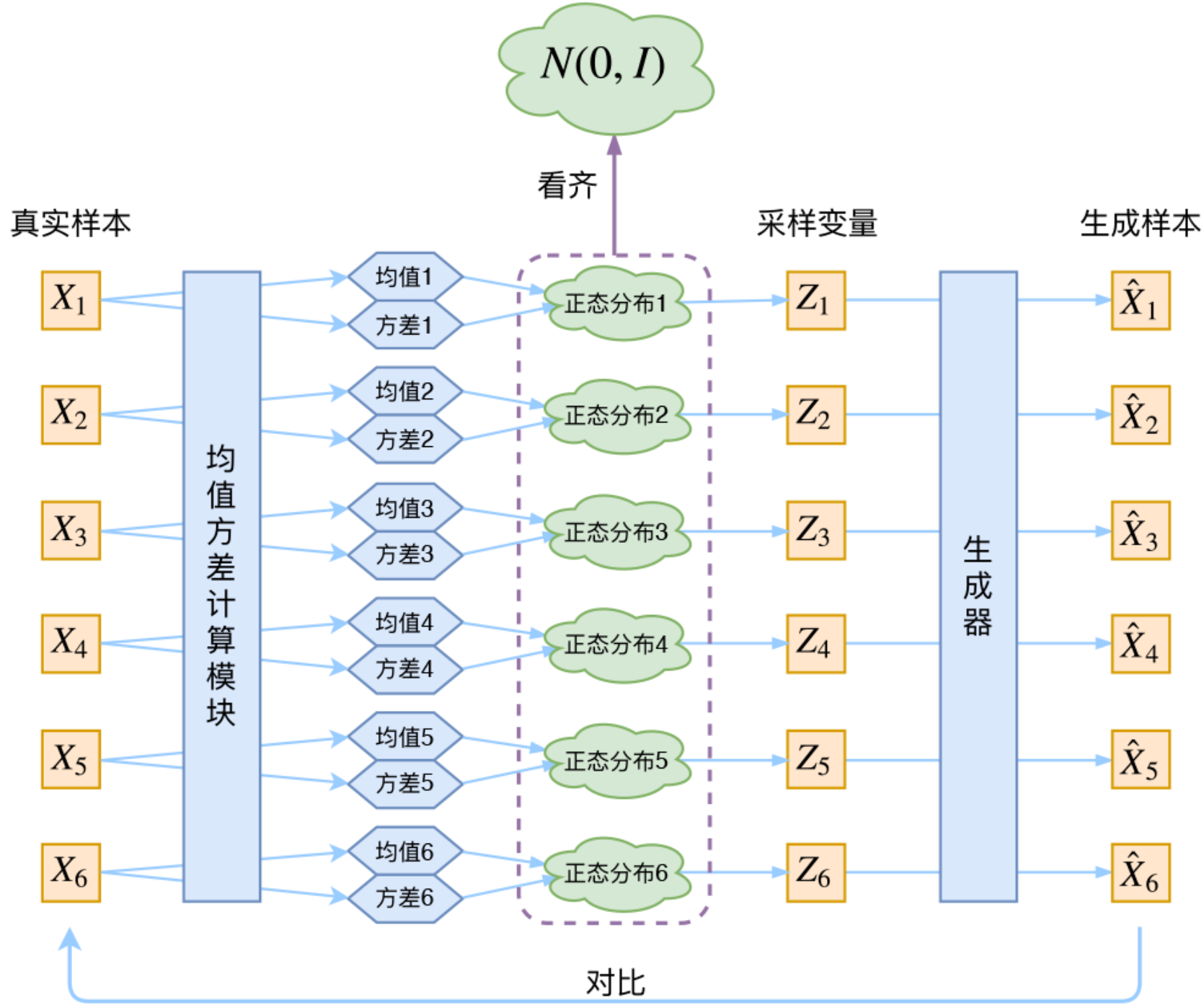


第二个问题的答案：用采样代替全部计算，用采样数量代替加权。

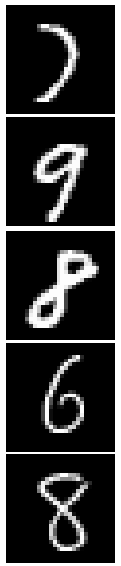






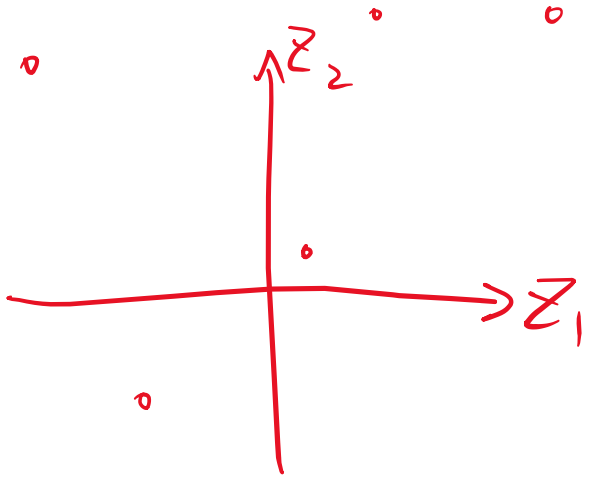

$N(0, I)$
看齐
真实样本
采样变量
生成样本
X_1
X_2
X_3
X_4
X_5
X_6
均值方差计算模块
均值1
方差1
均值2
方差2
均值3
方差3
均值4
方差4
均值5
方差5
均值6
方差6
正态分布1
正态分布2
正态分布3
正态分布4
正态分布5
正态分布6
Z_1
Z_2
Z_3
Z_4
Z_5
Z_6
生成器
$\hat{X}_1$
$\hat{X}_2$
$\hat{X}_3$
$\hat{X}_4$
$\hat{X}_5$
$\hat{X}_6$
对比

但是！依然会过拟合，而且跟AE一模一样！

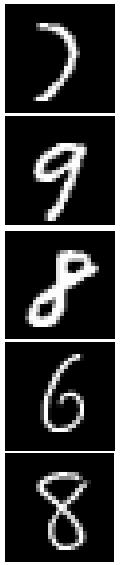


Encoder

[1,1, 0.003,0.002]
[-8,-8, 0.01,0.01]
[999,999, 0.001,0.0001]
[3.656,666,0.002,0.001]
[-123,765, 0.003,0.0002]



Decoder
d([1,1,*,*])=
d([8,8,*,*])=
d([999,999,*,*])=
d([3.656,666,*,*])=
d([123,765,*,*])=



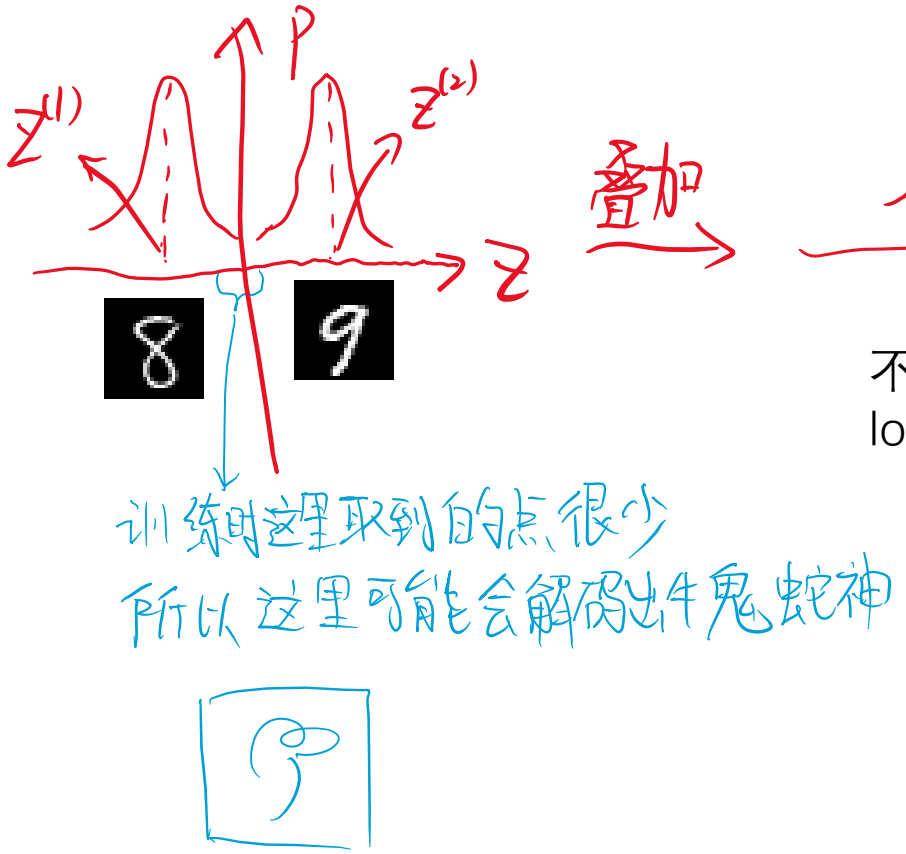
loss=0!

我们希望圈胖胖的，重叠越多越好

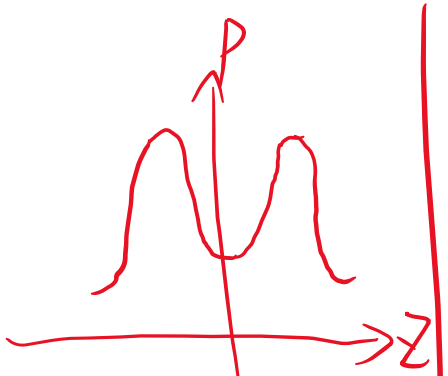
如何让圈胖胖的叠起来？

解决方法：加loss函数，让样本集的图片在隐空间的高斯分布圈圈们叠起来越像标准高斯分布越好。

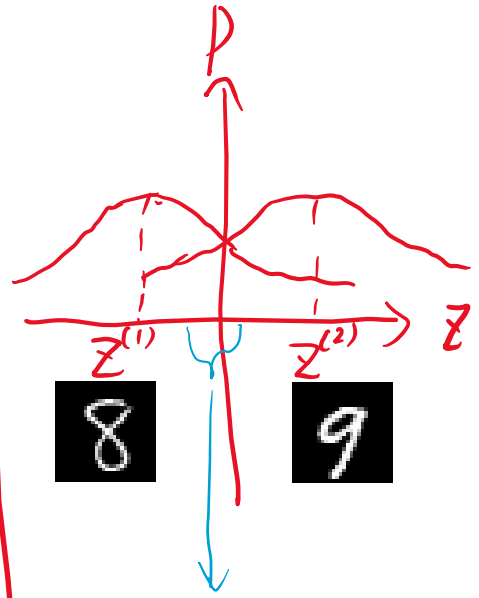
以一维高斯为例：



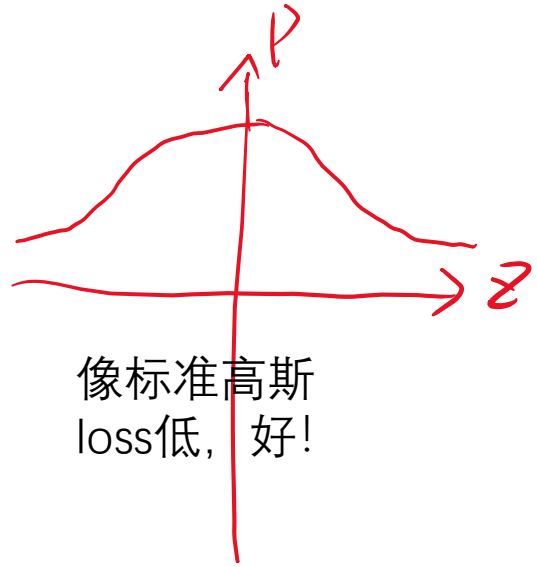
叠加



不像标准高斯
loss高，不好！



训练时，这里的点常被抓去算loss，因此它们必须学会圆滑处世，又像8又像9才行



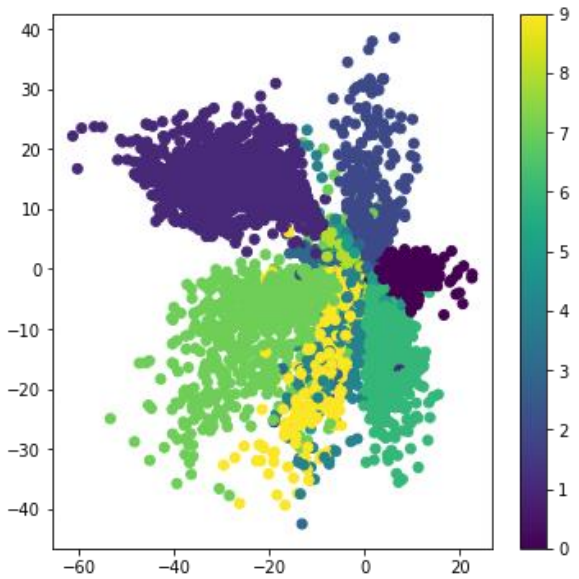
像标准高斯
loss低，好！

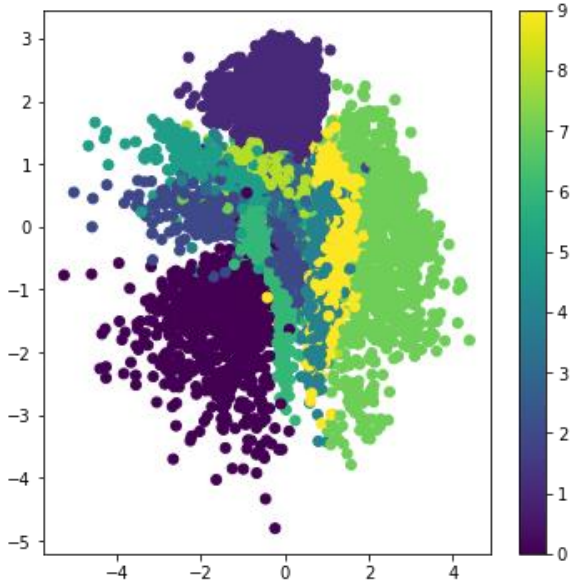
目的都达到了

通过以上操作，我们达成了：

- 隐空间有规律可循，长的像的图片离得近
- 隐空间随便拿个点解码之后，得到的点有意义。因为，除了图象被编码后的那些高斯中心点以外，每个没对应样本的点都有可能被抓去算loss，所以他们都要跟输入的图片长得像点才行。
- 换句话说，隐空间中对应不同标签的点不会离得很远（因为过渡点要被抓去算loss），但也不会离得太近（因为每个高斯的中心部分因为被采样次数多必须特色鲜明，不能跟别的类别的高斯中心离得太近）（VAE做生成任务的基础）
- 隐空间对应相同标签的点离得比较近，但又不会聚成超小的小簇，然而也不会有相聚甚远的情况（VAE做分类任务的基础）

AE

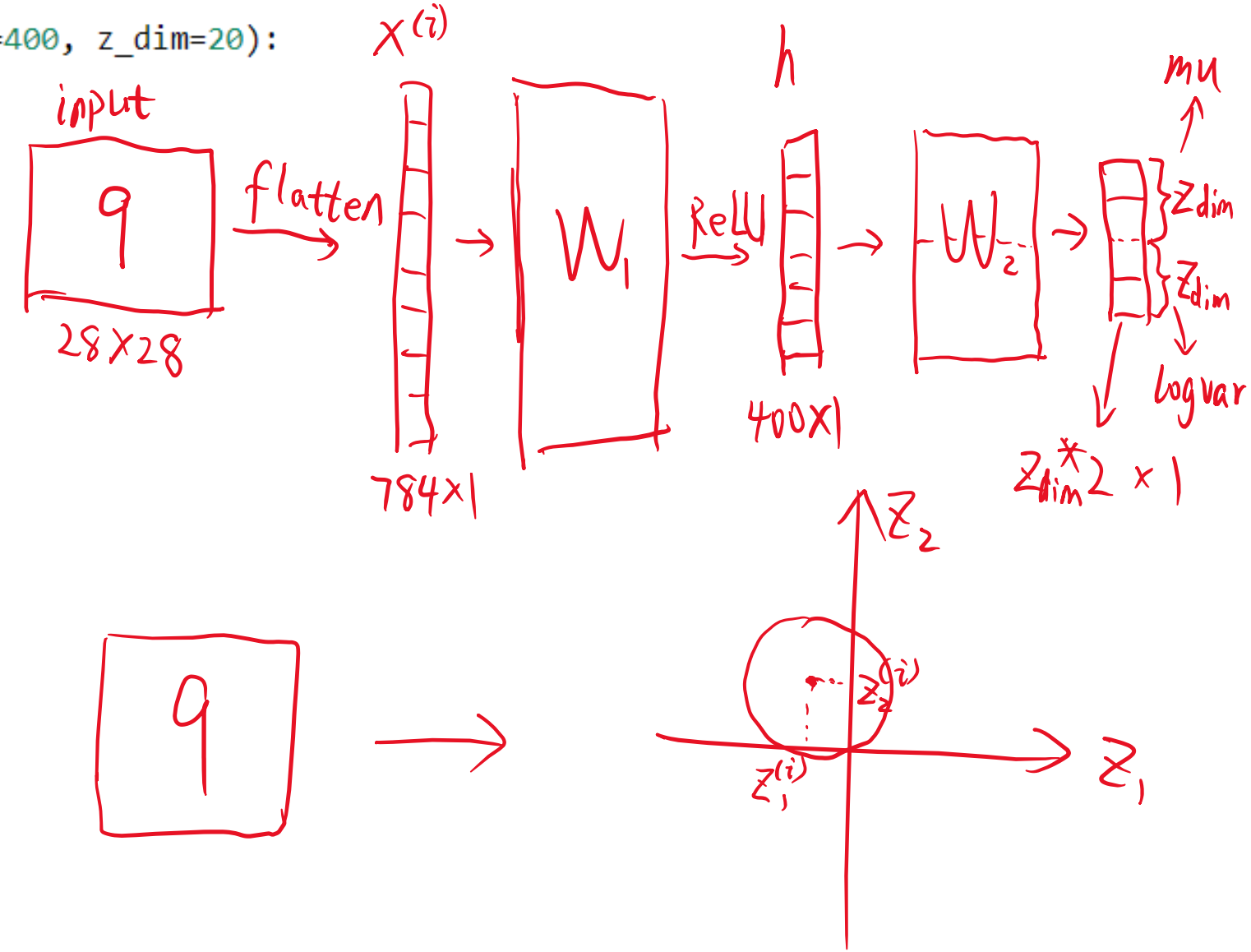




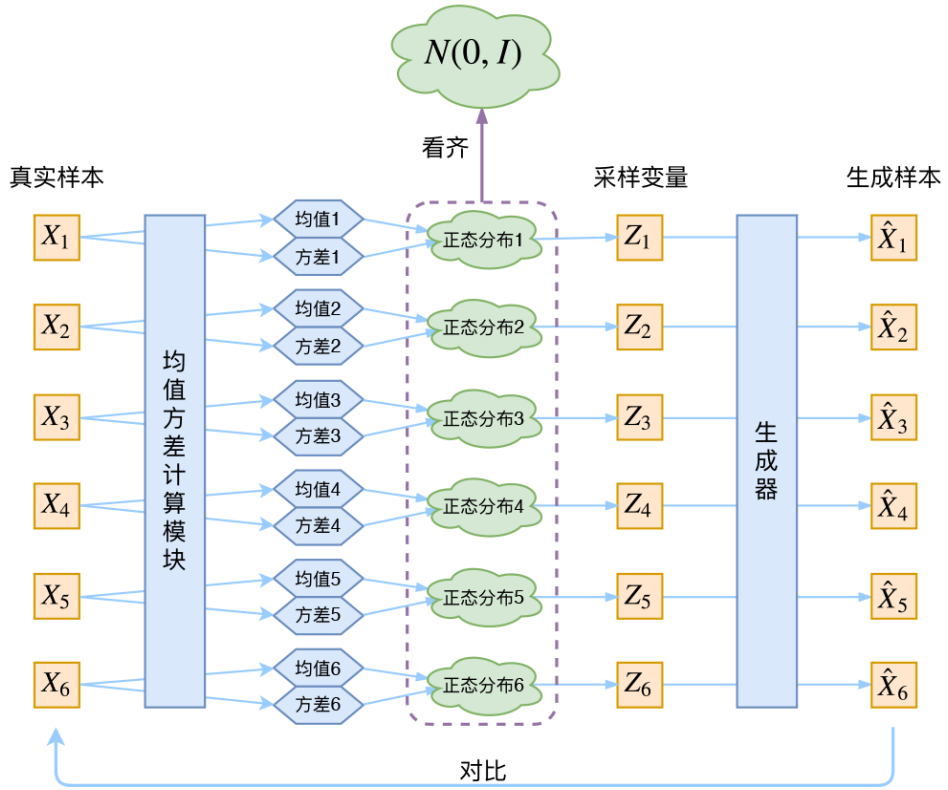
VAE

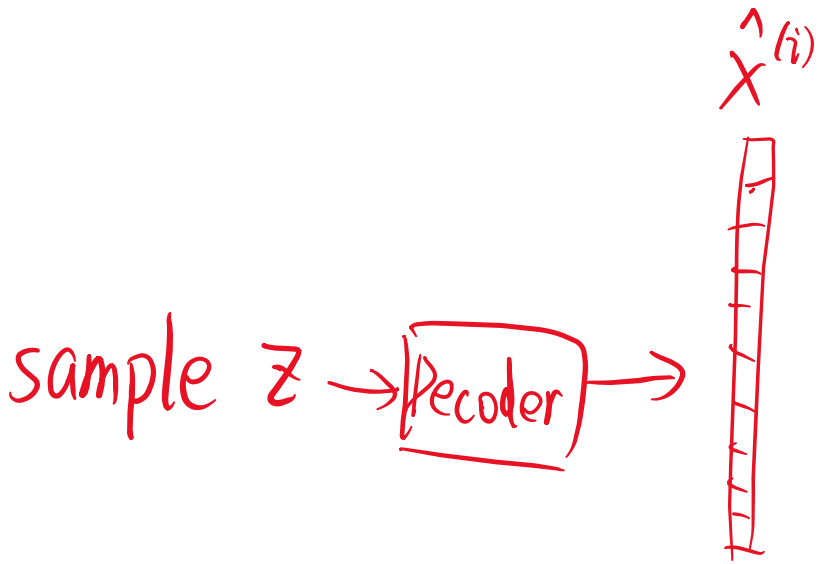
代码部分

```python
class VAE(nn.Module):
    def __init__(self, image_size=784, h_dim=400, z_dim=20):
        super(VAE, self).__init__()
        self.encoder = nn.Sequential(
            nn.Linear(image_size, h_dim),
            nn.LeakyReLU(0.2),
            nn.Linear(h_dim, z_dim*2)
        )
        self.decoder = nn.Sequential(
            nn.Linear(z_dim, h_dim),
            nn.ReLU(),
            nn.Linear(h_dim, image_size),
            nn.Sigmoid()
        )

    def reparameterize(self, mu, logvar):
        std = logvar.mul(0.5).exp_()
        esp = torch.randn(*mu.size())
        z = mu + std * esp
        return z

    def forward(self, x):
        h = self.encoder(x)
        mu, logvar = torch.chunk(h, 2, dim=-1)
        z = self.reparameterize(mu, logvar)
        return self.decoder(z), mu, logvar
```

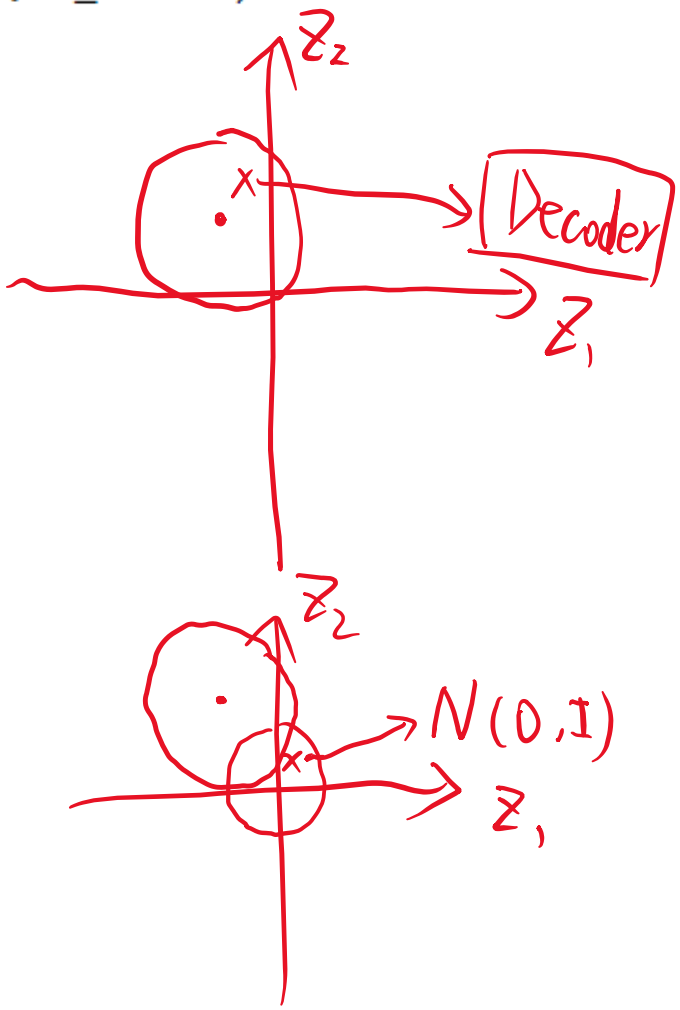


```python
class VAE(nn.Module):
    def __init__(self, image_size=784, h_dim=400, z_dim=20):
        super(VAE, self).__init__()
        self.encoder = nn.Sequential(
            nn.Linear(image_size, h_dim),
            nn.LeakyReLU(0.2),
            nn.Linear(h_dim, z_dim*2)
        )
        self.decoder = nn.Sequential(
            nn.Linear(z_dim, h_dim),
            nn.ReLU(),
            nn.Linear(h_dim, image_size),
            nn.Sigmoid()
        )

    def reparameterize(self, mu, logvar):
        std = logvar.mul(0.5).exp_()
        esp = torch.randn(*mu.size())
        z = mu + std * esp
        return z

    def forward(self, x):
        h = self.encoder(x)
        mu, logvar = torch.chunk(h, 2, dim=-1)
        z = self.reparameterize(mu, logvar)
        return self.decoder(z), mu, logvar
```





```python
class VAE(nn.Module):
    def __init__(self, image_size=784, h_dim=400, z_dim=20):
        super(VAE, self).__init__()
        self.encoder = nn.Sequential(
            nn.Linear(image_size, h_dim),
            nn.LeakyReLU(0.2),
            nn.Linear(h_dim, z_dim*2)
        )
        self.decoder = nn.Sequential(
            nn.Linear(z_dim, h_dim),
            nn.ReLU(),
            nn.Linear(h_dim, image_size),
            nn.Sigmoid()
        )

    def reparameterize(self, mu, logvar):
        std = logvar.mul(0.5).exp_()
        esp = torch.randn(*mu.size())
        z = mu + std * esp
        return z

    def forward(self, x):
        h = self.encoder(x)
        mu, logvar = torch.chunk(h, 2, dim=-1)
        z = self.reparameterize(mu, logvar)
        return self.decoder(z), mu, logvar
```



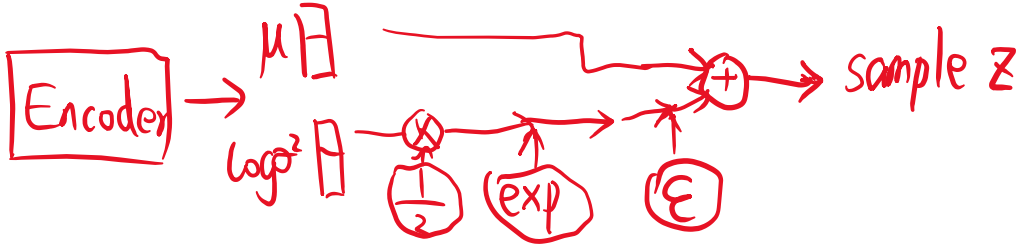
bad!

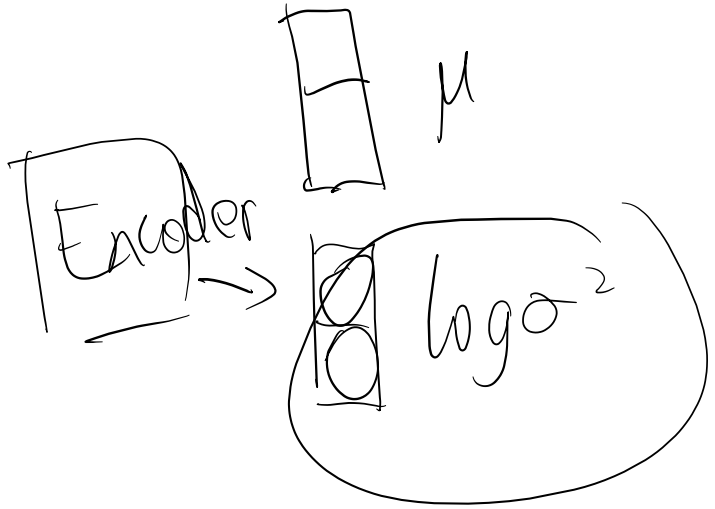
梯度断裂

$$\frac{N(\mu,\Sigma)-\mu}{\Sigma}=N(0,I)$$

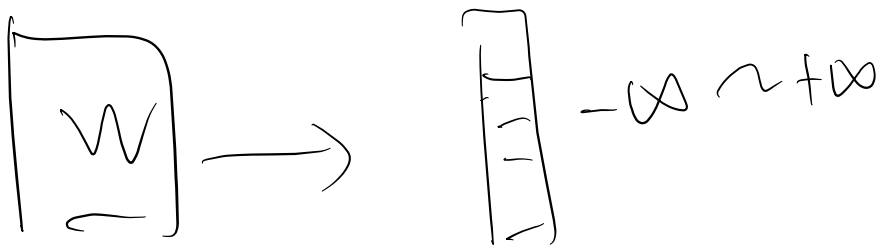
从$N(0,I)$中采样：ε

则$N(\mu,\Sigma)$中的样本可表示为：$\varepsilon\Sigma+\mu$





$$\frac{1}{2}\log\sigma^2 = \log(\sigma^2)^{\frac{1}{2}} = \log\sigma$$



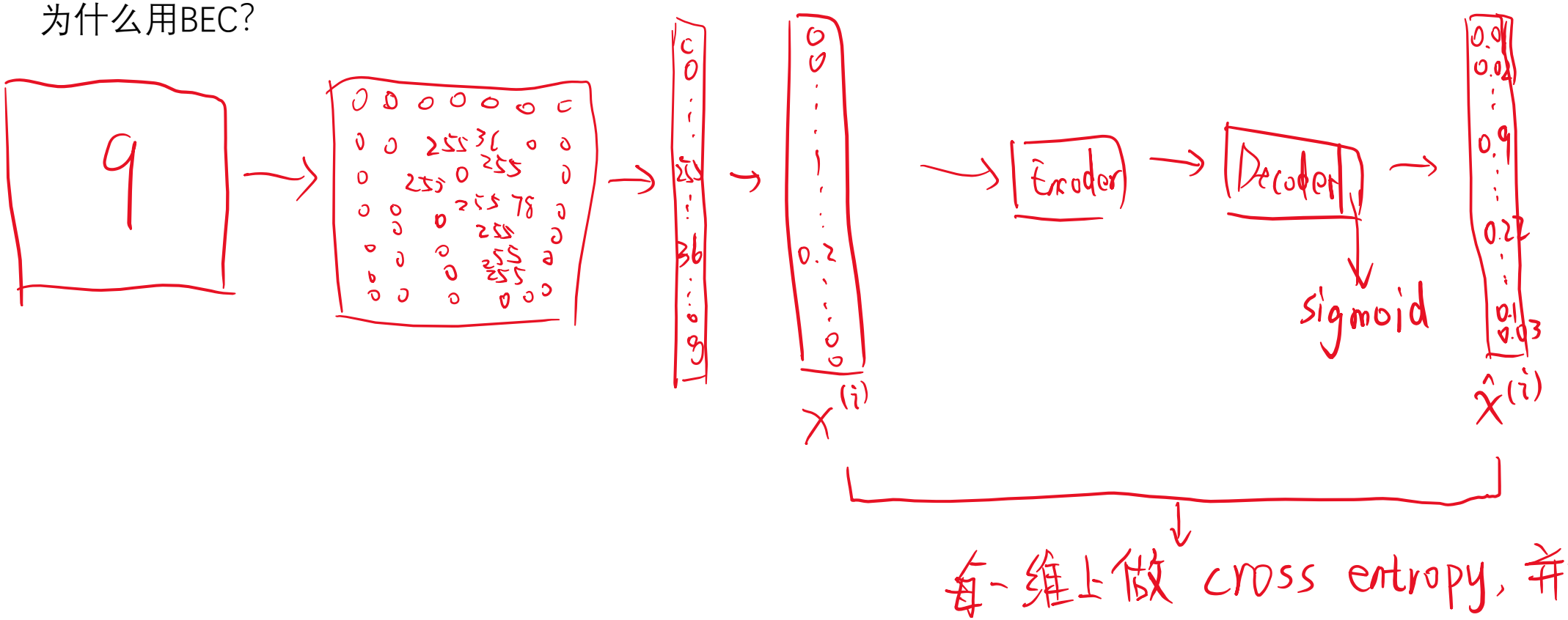
$$e^{\log\sigma} = \sigma$$

$$\epsilon \cdot \sigma + \mu$$

```python
def loss_fn(recon_x, x, mu, logvar, beta):
    BEC = F.binary_cross_entropy(recon_x, x, size_average=False)
    KLD = -0.5 * torch.sum(1 + logvar - mu**2 - logvar.exp())
    return BCE + beta * KLD
```

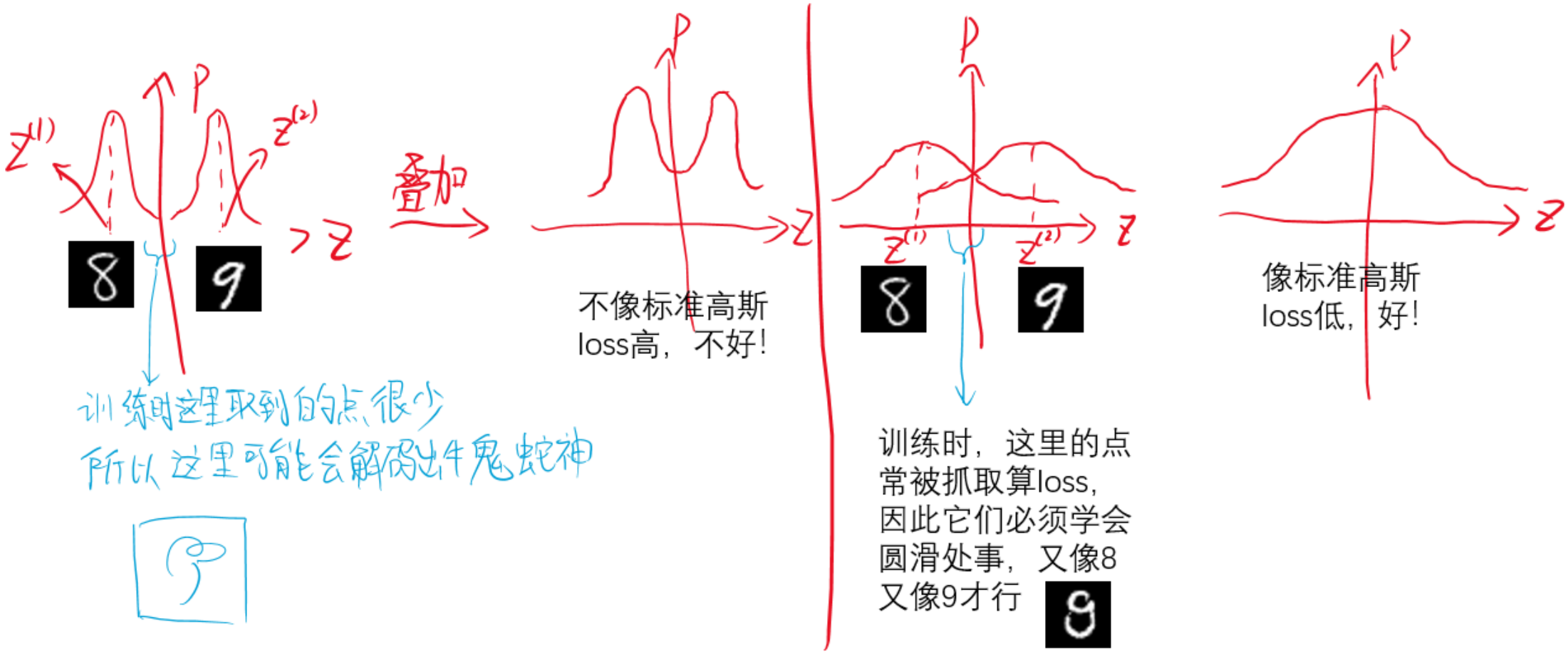
$$BEC = binary\_cross\_entropy = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(p(y_i)) + (1 - y_i) \cdot log(1 - p(y_i)) \quad \times N$$

为什么用BEC?



```python
def loss_fn(recon_x, x, mu, logvar, beta):
    BEC = F.binary_cross_entropy(recon_x, x, size_average=False)
    KLD = -0.5 * torch.sum(1 + logvar - mu**2 - logvar.exp())
    return BCE + beta * KLD
```

KLD（KL散度, KL Divergence，相对熵）：要让所有样本的隐变量圈圈叠加成标准高斯分布N(0,I)



KLD的公式是怎么来的呢？见数学部分的第三子部分

数学部分

1. 熵、交叉熵、KL散度的概念

2.VAE loss函数的由来

```
return BCE + beta * KLD
```

3.VAE KLD部分公式的推导

```
KLD = -0.5 * torch.sum(1 + logvar - mu**2 - logvar.exp())
```

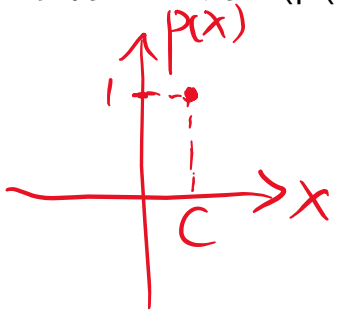
数学部分

1. 熵、交叉熵、KL散度的概念

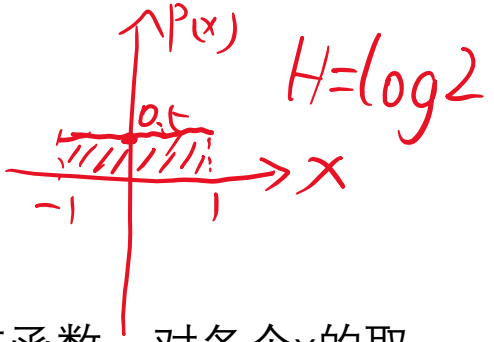
熵（Entropy）

假设p(x)是一个分布函数，满足在x上的积分为1，那么p(x)的熵定义为H(p(x))，这里我们简写为H(p)

$$H(p) = \int p(x) \log \frac{1}{p(x)} dx$$



∴ H=0



H=log2

直观上，越分散的分布函数熵越大。越集中的分布函数熵越小。熵的最小值为0.

∵ $p(x) \in [0,1]$，$\frac{1}{p(x)} \geq 1$

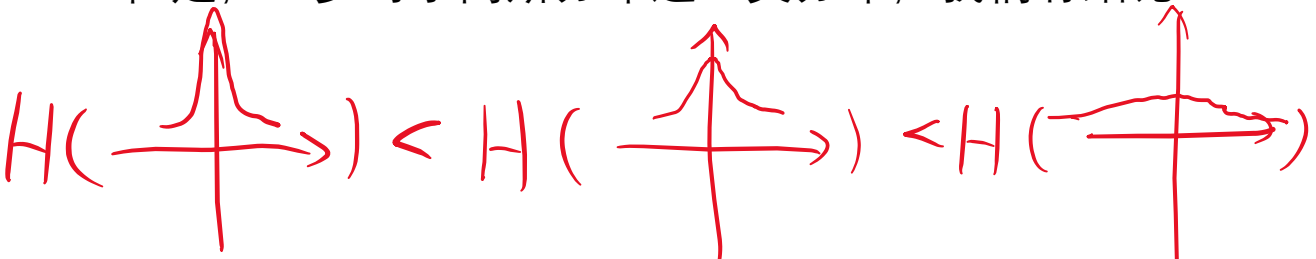
∴ $p(x) \log \frac{1}{p(x)} \geq 0$

∴ $\int p(x) \log \frac{1}{p(x)} dx \geq 0$

当 $p(C)=1$ 时（C为常数）

$H(p) = p(C) \log \frac{1}{p(C)} = 1 \cdot \log \frac{1}{1} = 1 \cdot 0 = 0$

事实上，对于一般的、分散的分布函数，对各个x的取值，log(1/p(x))会变大，可参与作和的x会变多，而p(x)会变小，所以整体算完H(p)不见得比集中的分布函数大，但是，至少对于高斯分布这一类分布，我们有结论

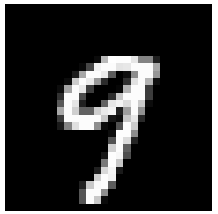


从信息论的角度来说，熵又叫信息熵，它的大小表示信息量的多少，分散的分布函数可能性多、拿到p(x)后对于x的推断不确定性大，即信息量大，而对于p(C)=1这种情况，拿到分布函数直接就拿到了结果，因此信息量为0

交叉熵（Cross-Entropy）

假设p(x)、q(x)是两个分布函数，交叉熵的小大评价了这两个分布函数的相似与否。p和q的交叉熵记为H(p, q)
交叉熵小—分布相似；交叉熵大—分布不相似。

$$H(p,q)=\int p(x)\log\frac{1}{q(x)}dx$$ ，注意，$H(p,q)$ 不一定等于 $H(q,p)$



$x^{(i)}$

→ 分类器（softmax）→

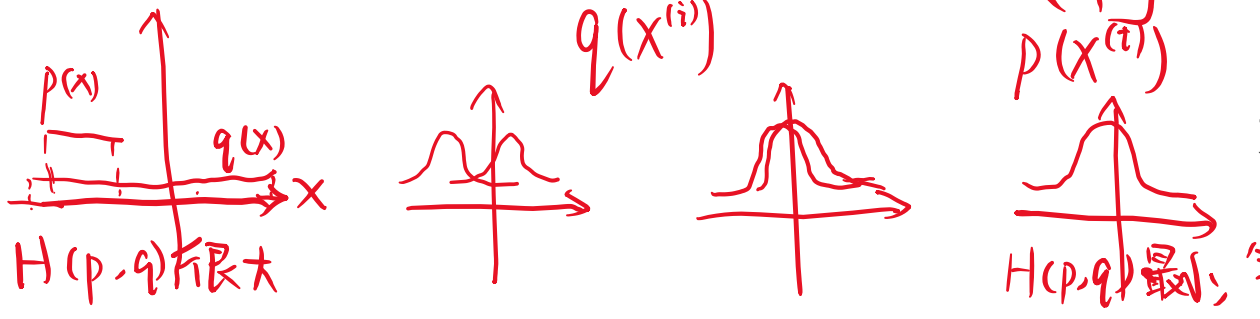
$q(x^{(i)})$: [0.01, 0.02, 0.01, 0.05, 0.04, 0.01, 0.01, 0.01, 0.01, 0.83]

⟷

$p(x^{(i)})$: [0, 0, 0, 0, 0, 0, 0, 0, 0, 1]

$$H(p(x^{(i)}),q(x^{(i)}))=0+0+\cdots+0+1\cdot\log\frac{1}{0.83}=\log\frac{1}{0.83}$$

对于连续分布：



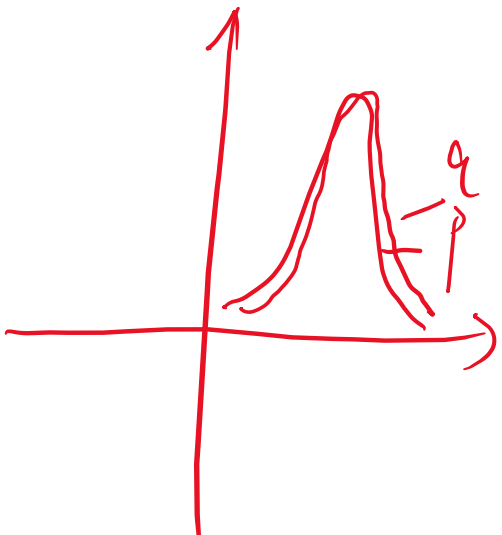
交叉熵最大为非常大，最小为p的熵H(p)

KL散度

假设p(x)、q(x)是两个分布函数，KL散度的小大评价了这两个分布函数的相似与否，同时考虑了p(x)这个分布的信息量。记为KL(p, q)。注意：KL(p, q)也不一定等于KL(q,p)。

$$KL(p,q) = H(p,q) - H(p)$$

KL散度小---分布相似 | [p(x)分散 | p(x)信息量大]。
KL散度大---分布不相似 | [p(x)集中 | p(x)信息量小]。
KL散度最小值为0（因为H(p,q)最小值为H(p)）： p(x)和q(x)完全相同时。



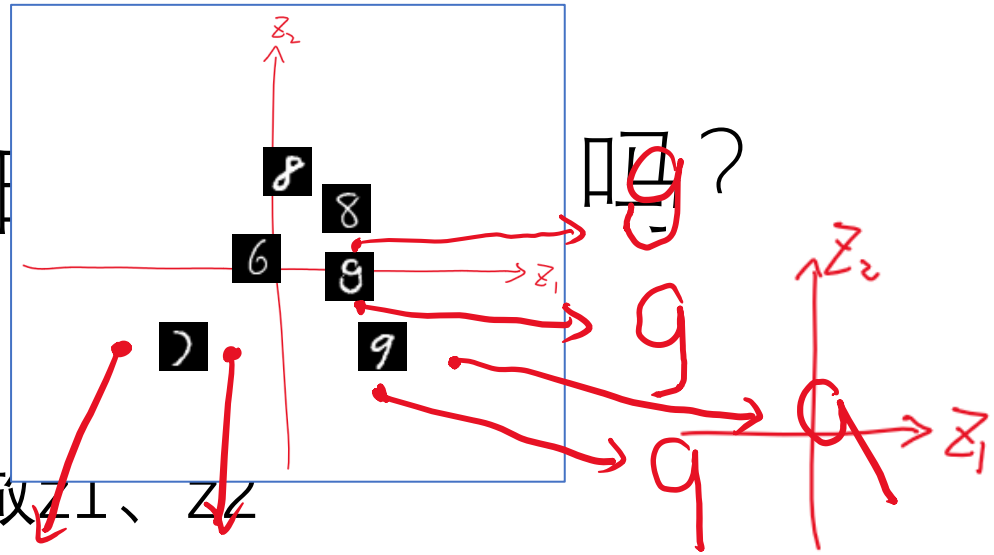
H(p)很小，-H(p)很大
但H(p,q)很接近其最小值H(p)
故KL(p,q)很小

H(p,q)很大
但H(p)很大
-H(p)很小
故KL(p,q)很小

数学部分

2.VAE loss函数的由来

MNIST手写数字数据集真[illegible]吗？

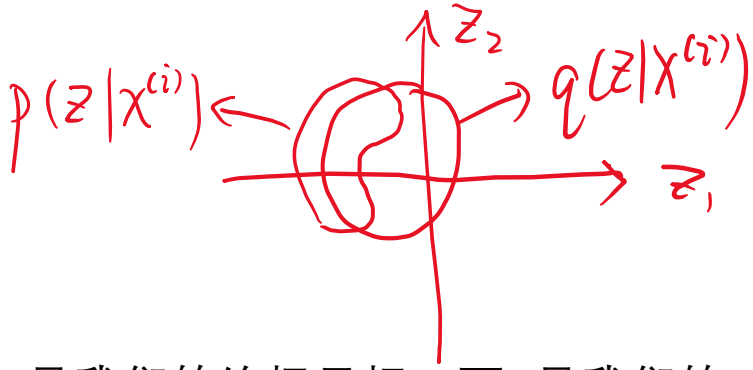


- 不!
- 很久很久以前，有一个二维平面，叫做z1、z2
- 有一位全知全能的神，将过去、现在与未来的、世界上无穷的手写数字都井然有序地映射到了这个二维平面上
- 因为有无穷个手写数字，这个二维平面上的任何一个点都是有意义的，这个平面就是我们VAE模型追求的究极平面。
- 神在它的笔记本上记下了p(z)和p(x|z)这两个分布函数，并宣称：以后人类所写的所有手写数字，将从这两个分布中产生。
- 你可以将p(z)理解为，当你想写一个数字时，你脑海中数字的样子，而p(x|z)则是你实际上写下这个数字时，因为风吹草动写歪来的、跟你脑袋里想的那个有点不一样的数字。

MNIST手写数字数据集真的是人写的吗？

- 神编的隐变量z是好的，因为它低维，且饱含意义，人类好想把图片编成神所编成的z啊
- 现在人们想探求神是如何把一切图片映射到二维平面上的，也就是想知道神做映射时候用的p(z|x)。
- 根据贝叶斯公式，$\mathrm{p}(z|x) = \frac{p(x|z)p(z)}{p(x)} = \frac{p(x|z)p(z)}{\iint_z p(x|z)p(z)dz}$
- 好家伙，p(x|z)、 p(z|x)和p(z)都是神的小秘密，我上哪知道去
- 索性统统丢掉，让无敌的神经网络和梯度下降来近似这一切吧!

VAE的loss函数

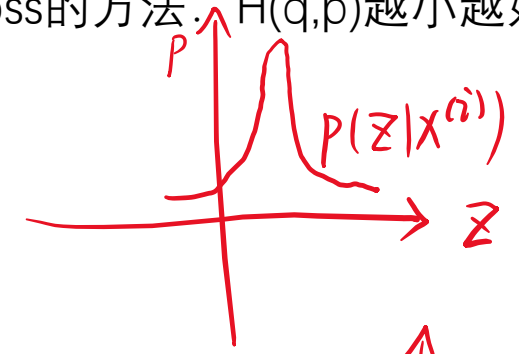


L=KL(q(z|x), p(z|x))，其中p(z|x)是神的分布，是我们的终极目标，而q是我们的Encoder，我们用Encoder算出来的高斯分布的均值和方差去近似神的分布。

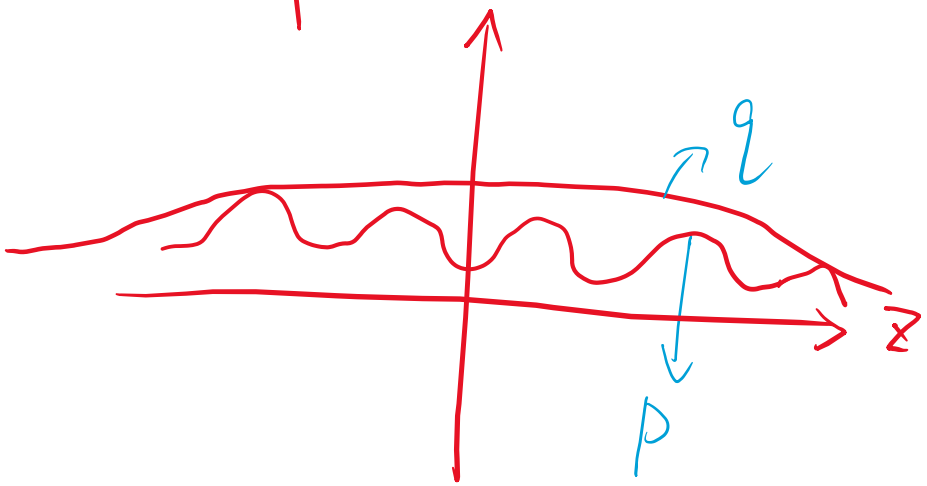
$$Loss = KL(q,p) = H(q,p) - H(q)$$

降低loss的方法：H(q,p)越小越好，H(q)越大越好

如果神对于某张图片的编码长得像一个熵很大的高斯，那我们就不用担心H(q)很小，只要能好好套住它，这张图片的loss也能降到0



如果神对于某张图片的编码分布长得不像高斯，那我们也没法硬套，只好让H(q)大点，也就是我们的高斯扁平一些，来尽可能地近似神的编码分布



然而，p(z|x)我们也不知道，那咋算KLD啊？

- 这回有trick!

利用 $\int p(x)\log\frac{1}{q(x)}dx = E_{p(x)}(\log\frac{1}{q(x)})$

$H(p) = E_p(\log\frac{1}{p}) = -E_p(\log p)$

$H(p,q) = E_p(\log\frac{1}{q}) = -E_p(\log q)$

$$Loss = KL(q(z|x^{(i)}), p(z|x^{(i)}))$$

$$= H(q(z|x^{(i)}), p(z|x^{(i)})) - H(q(z|x^{(i)}))$$

$$= [-E_{q(z|x^{(i)})}(\log\frac{p(z)p(x^{(i)}|z)}{p(x^{(i)})})] - [-E_{q(z|x^{(i)})}(\log q(z|x^{(i)}))]$$

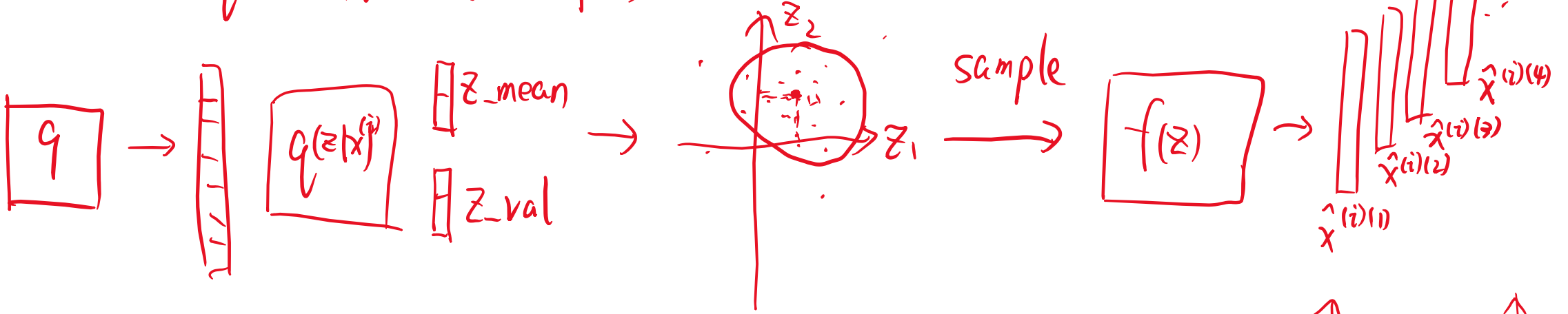
$$= \underbrace{-E_{q(z|x^{(i)})}(\log p(z))}_{H(q(z|x^{(i)}), p(z))} - E_{q(z|x^{(i)})}(\log p(x^{(i)}|z)) + \underbrace{E_{q(z|x^{(i)})}(\log p(x^{(i)}))}_{\log p(x^{(i)})\int q(z|x^{(i)})dz} + \underbrace{E_{q(z|x^{(i)})}(\log q(z|x^{(i)}))}_{-H(q(z|x^{(i)}))}$$

$$= -E_{q(z|x^{(i)})}(\log p(x^{(i)}|z)) + KL(q(z|x^{(i)}), p(z)) + \log p(x^{(i)})$$

$$Loss = -E_{q(z|x^{(i)})}(\log p(x^{(i)}|z)) + KL(q(z|x^{(i)}), p(z))$$

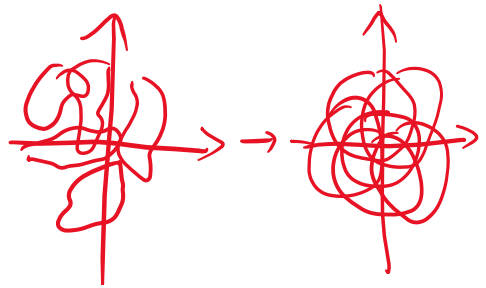
$$= E_{q(z|x^{(i)})}\left(\frac{\|x^{(i)} - f(z)\|^2}{2c}\right) + KL(q(z|x^{(i)}), p(z))$$

$q(z|x^{(i)})$ 为 Encoder，$f(z)$ 为 decoder



训练时每一个Epoch出下面的一条

loss函数里剩下的神的秘密，就只剩下p(z)了
这里就让我们猜一下神的小心思吧，比如代码里用的，世界上最好的分布——标准高斯分布



c是常数，表示我们更愿意相信手里的数据集，还是更愿意相信我们选择的先验分布

$$Loss = KL(q(z|x^{(i)}), p(z|x^{(i)}))$$

我们从希望编码器对每一个样本的分布接近对应的神的分布，变成了希望编码器对每一个样本的分布都接近先验分布（标准高斯），同时又要满足重构误差尽量小，了

从另一个角度看loss

$$Loss = KL(q(z|x^{(i)}), p(z|x^{(i)}))$$

$$= -E_{q(z|x^{(i)})}(\log p(x^{(i)}|z)) + KL(q(z|x^{(i)}), p(z)) + \log p(x^{(i)})$$

$$\geq 0$$

即 $$\log p(x^{(i)}) \geq \underbrace{E_{q(z|x^{(i)})}(\log p(x^{(i)}|z)) - KL(q(z|x^{(i)}), p(z))}_{\text{Evidence Lower Bound (ELBo)}}$$

要最大化 $\log p(x^{(i)})$，最大化ELBo即可

数学部分

3.VAE KLD部分公式的推导

$$KL(q(z|x), p(z))$$

```
KLD = -0.5 * torch.sum(1 + logvar - mu**2 - logvar.exp())
```

以下公式中请把所有的x都当成z

$$p(z) \to \frac{1}{\sqrt{2\pi\sigma_p^2}} \exp\left(-\frac{(z-\mu_p)^2}{2\sigma_p^2}\right)$$

$$q_\theta(z|x_i) \to \frac{1}{\sqrt{2\pi\sigma_q^2}} \exp\left(-\frac{(z-\mu_q)^2}{2\sigma_q^2}\right)$$

μ_p, σ_p 是先验分布的参数，是常数

μ_q, σ_q 是 Encoder 算出来的，在单个样本的 loss 计算中也是常数

$$-D_{KL}(q_\theta(z|x_i)||p(z)) =$$

$$\int \frac{1}{\sqrt{2\pi\sigma_q^2}} \exp\left(-\frac{(x-\mu_q)^2}{2\sigma_q^2}\right) \log\left(\frac{\frac{1}{\sqrt{2\pi\sigma_p^2}} \exp\left(-\frac{(x-\mu_p)^2}{2\sigma_p^2}\right)}{\frac{1}{\sqrt{2\pi\sigma_q^2}} \exp\left(-\frac{(x-\mu_q)^2}{2\sigma_q^2}\right)}\right) dz$$

$$= \int \frac{1}{\sqrt{2\pi\sigma_q^2}} \exp\left(-\frac{(x-\mu_q)^2}{2\sigma_q^2}\right) \times \left\{-\frac{1}{2}\log(2\pi) - \log(\sigma_p) - \frac{(x-\mu_p)^2}{2\sigma_p^2} + \frac{1}{2}\log(2\pi) + \log(\sigma_q) + \frac{(x-\mu_q)^2}{2\sigma_q^2}\right\} dz.$$

$$= \frac{1}{\sqrt{2\pi\sigma_q^2}} \int \exp\left(-\frac{(x-\mu_q)^2}{2\sigma_q^2}\right) \left\{-\log(\sigma_p) - \frac{(x-\mu_p)^2}{2\sigma_p^2} + \log(\sigma_q) + \frac{(x-\mu_q)^2}{2\sigma_q^2}\right\} dz.$$

$$= \frac{1}{\sqrt{2\pi\sigma_q^2}} \int \exp\left(-\frac{(x-\mu_q)^2}{2\sigma_q^2}\right) \left\{\log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{(x-\mu_p)^2}{2\sigma_p^2} + \frac{(x-\mu_q)^2}{2\sigma_q^2}\right\} dz.$$

$$
\begin{aligned}
-D_{KL}(q_\theta(z|x_i)||p(z)) &= E_q\left\{\log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{(x-\mu_p)^2}{2\sigma_p^2} + \frac{(x-\mu_q)^2}{2\sigma_q^2}\right\} \\
&= \log\left(\frac{\sigma_q}{\sigma_p}\right) + E_q\left\{-\frac{(x-\mu_p)^2}{2\sigma_p^2} + \frac{(x-\mu_q)^2}{2\sigma_q^2}\right\} \\
&= \log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{1}{2\sigma_p^2}E_q\left\{(x-\mu_p)^2\right\} + \frac{1}{2\sigma_q^2}E_q\left\{(x-\mu_q)^2\right\}
\end{aligned}
$$

$$
\sigma_q^2 = E_q\left\{(x-\mu_q)^2\right\}
$$

$$
\begin{aligned}
-D_{KL}(q_\theta(z|x_i)||p(z)) &= \log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{1}{2\sigma_p^2}E_q\left\{(x-\mu_p)^2\right\} + \frac{\sigma_q^2}{2\sigma_q^2} \\
&= \log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{1}{2\sigma_p^2}E_q\left\{(x-\mu_p)^2\right\} + \frac{1}{2} \\
&= \log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{1}{2\sigma_p^2}E_q\left\{(x-\mu_q+\mu_q-\mu_p)^2\right\} + \frac{1}{2} \\
&= \log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{1}{2\sigma_p^2}E_q\left\{(\underbrace{x-\mu_q}_{a} + \underbrace{\mu_q-\mu_p}_{b})^2\right\} + \frac{1}{2}
\end{aligned}
$$

$$= \log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{1}{2\sigma_p^2} E_q\left\{(x-\mu_q)^2 + 2(x-\mu_q)(\mu_q-\mu_p) + (\mu_q-\mu_p)^2\right\} + \frac{1}{2}$$

$$= \log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{1}{2\sigma_p^2} E_q\left\{(x-\mu_q)^2 + 2(x-\mu_q)(\mu_q-\mu_p) + (\mu_q-\mu_p)^2\right\} + \frac{1}{2}$$

$$= \log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{1}{2\sigma_p^2}\left[E_q\left\{(x-\mu_q)^2\right\} + 2E_q\left\{(x-\mu_q)(\mu_q-\mu_p)\right\} + E_q\left\{(\mu_q-\mu_p)^2\right\}\right] + \frac{1}{2}$$

$$= \log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{1}{2\sigma_p^2}\left[\sigma_q^2 + 2 * 0 * (\mu_q-\mu_p) + (\mu_q-\mu_p)^2\right] + \frac{1}{2}$$

$$= \log\left(\frac{\sigma_q}{\sigma_p}\right) - \frac{\sigma_q^2 + (\mu_q-\mu_p)^2}{2\sigma_p^2} + \frac{1}{2}$$

$\hat{\nu}$ $\mu_p = \vec{0}$, $\sigma_p = \vec{1}$

$$-D_{KL}(q_\theta(z|x_i)||p(z)) = \log(\sigma_q) - \frac{\sigma_q^2 + \mu_q^2}{2} + \frac{1}{2}$$

$$= \frac{1}{2}\log(\sigma_q^2) - \frac{\sigma_q^2 + \mu_q^2}{2} + \frac{1}{2}$$

$$= \frac{1}{2}\left[1 + \log(\sigma_q^2) - \sigma_q^2 - \mu_q^2\right]$$

$KL(q(z|x), p(z))$

```
KLD = -0.5 * torch.sum(1 + logvar - mu**2 - logvar.exp())
```

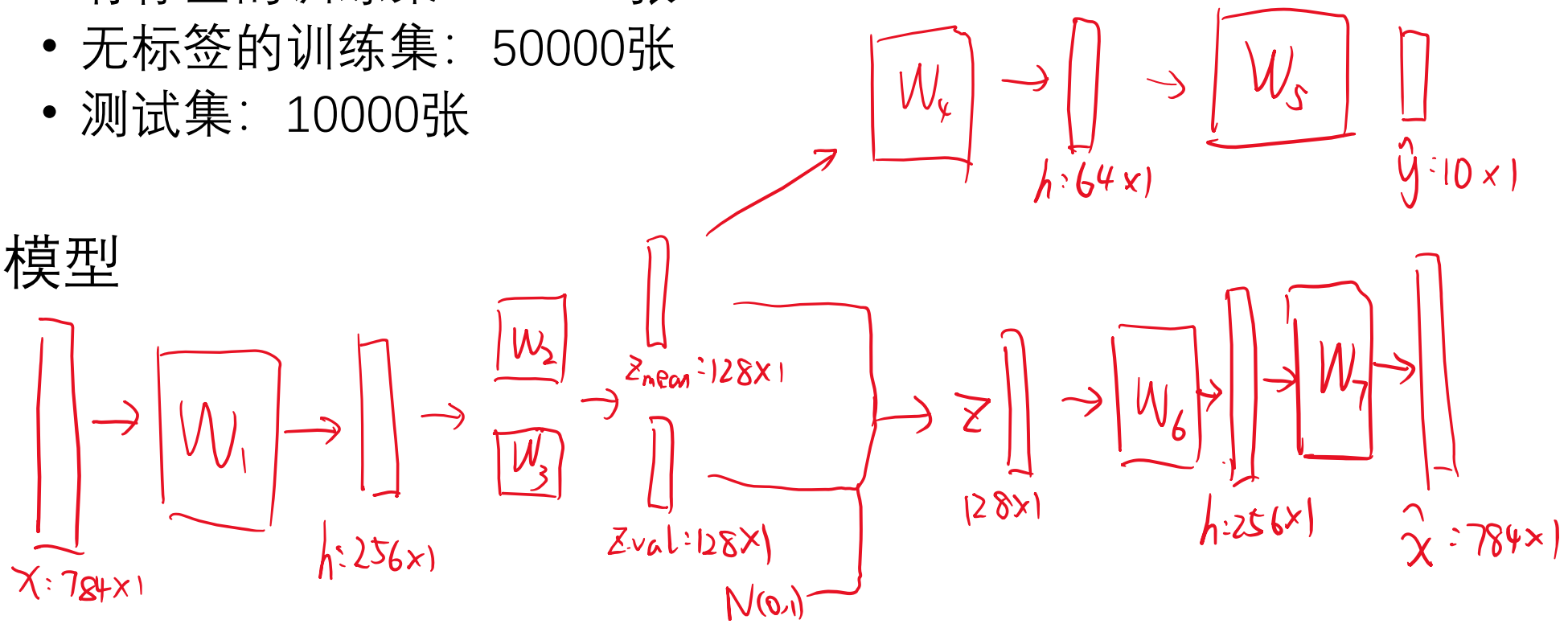
VAE辅助分类任务

半监督学习

数据集

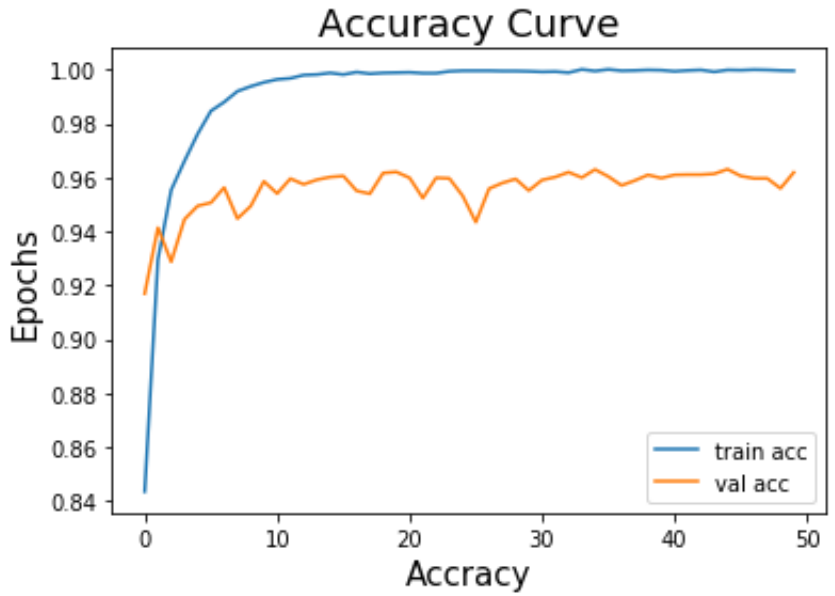
- MNIST：训练集60000张，测试集10000张
- 重新划分如下
  - 有标签的训练集：10000张
  - 无标签的训练集：50000张
  - 测试集：10000张

- 模型



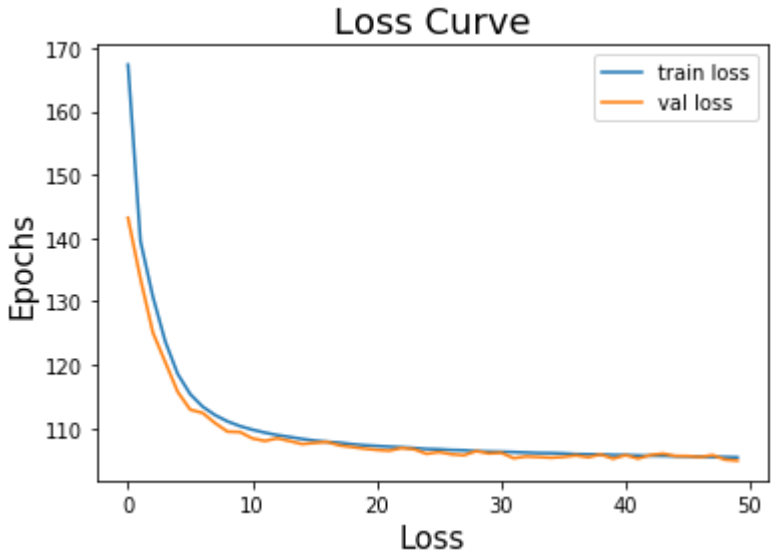
结果

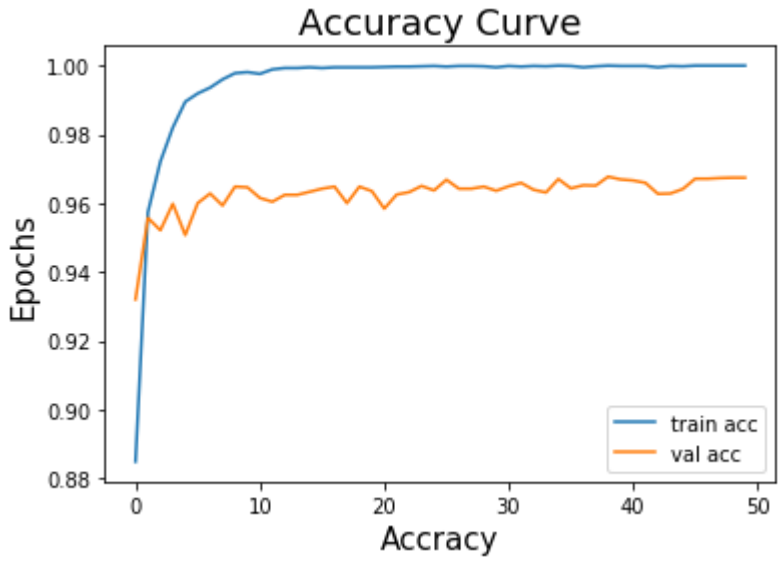
直接用带标签的10000张训练50轮



loss: 0.0011 - acc: 0.9997 - val_loss: 0.3403 - val_acc: 0.9602

先在60000张的训练集上跑50轮VAE，再用带标签的10000张训练50轮





loss: 105.4156 - val_loss: 104.8996

不加VAE:

```
Epoch 42/50
10000/10000 [==============================] - 1s 71us/step - loss: 3.4345e-04 - acc: 0.9999 - val_loss: 0.4154 - val_acc: 0.9542
Epoch 43/50
10000/10000 [==============================] - 1s 72us/step - loss: 4.6876e-04 - acc: 0.9998 - val_loss: 0.3393 - val_acc: 0.9616
Epoch 44/50
10000/10000 [==============================] - 1s 73us/step - loss: 4.5145e-05 - acc: 1.0000 - val_loss: 0.3718 - val_acc: 0.9588
Epoch 45/50
10000/10000 [==============================] - 1s 73us/step - loss: 0.0063 - acc: 0.9990 - val_loss: 0.3819 - val_acc: 0.9584
Epoch 46/50
10000/10000 [==============================] - 1s 71us/step - loss: 2.0943e-05 - acc: 1.0000 - val_loss: 0.3583 - val_acc: 0.9609
Epoch 47/50
10000/10000 [==============================] - 1s 70us/step - loss: 0.0032 - acc: 0.9992 - val_loss: 0.3544 - val_acc: 0.9606
Epoch 48/50
10000/10000 [==============================] - 1s 73us/step - loss: 0.0033 - acc: 0.9992 - val_loss: 0.3726 - val_acc: 0.9574
Epoch 49/50
10000/10000 [==============================] - 1s 70us/step - loss: 8.1487e-04 - acc: 0.9996 - val_loss: 0.3869 - val_acc: 0.9564
Epoch 50/50
10000/10000 [==============================] - 1s 71us/step - loss: 0.0011 - acc: 0.9997 - val_loss: 0.3403 - val_acc: 0.9602
```

加VAE:

```
Epoch 42/50
10000/10000 [==============================] - 1s 69us/step - loss: 6.4937e-04 - acc: 0.9998 - val_loss: 0.3644 - val_acc: 0.9618
Epoch 43/50
10000/10000 [==============================] - 1s 68us/step - loss: 0.0014 - acc: 0.9997 - val_loss: 0.3469 - val_acc: 0.9648
Epoch 44/50
10000/10000 [==============================] - 1s 69us/step - loss: 8.3939e-05 - acc: 1.0000 - val_loss: 0.3302 - val_acc: 0.9648
Epoch 45/50
10000/10000 [==============================] - 1s 69us/step - loss: 1.1915e-06 - acc: 1.0000 - val_loss: 0.3211 - val_acc: 0.9658
Epoch 46/50
10000/10000 [==============================] - 1s 73us/step - loss: 2.0397e-04 - acc: 0.9999 - val_loss: 0.3221 - val_acc: 0.9671
Epoch 47/50
10000/10000 [==============================] - 1s 69us/step - loss: 6.9571e-05 - acc: 1.0000 - val_loss: 0.3176 - val_acc: 0.9665
Epoch 48/50
10000/10000 [==============================] - 1s 68us/step - loss: 1.0319e-04 - acc: 0.9999 - val_loss: 0.3609 - val_acc: 0.9617
Epoch 49/50
10000/10000 [==============================] - 1s 73us/step - loss: 0.0012 - acc: 0.9998 - val_loss: 0.3240 - val_acc: 0.9669
Epoch 50/50
10000/10000 [==============================] - 1s 73us/step - loss: 1.6315e-07 - acc: 1.0000 - val_loss: 0.3127 - val_acc: 0.9668
```

VAE做文本生成

经典文章

ICLR2016

Generating Sentences from a Continuous Space

Samuel R. Bowman*
NLP Group and Dept. of Linguistics
Stanford University
sbowman@stanford.edu

Luke Vilnis*
CICS
UMass Amherst
luke@cs.umass.edu

Oriol Vinyals, Andrew M. Dai, Rafal Jozefowicz & Samy Bengio
Google Brain
Google, Inc.
{vinyals, adai, bengio}@google.com, rafjoz@gmail.com

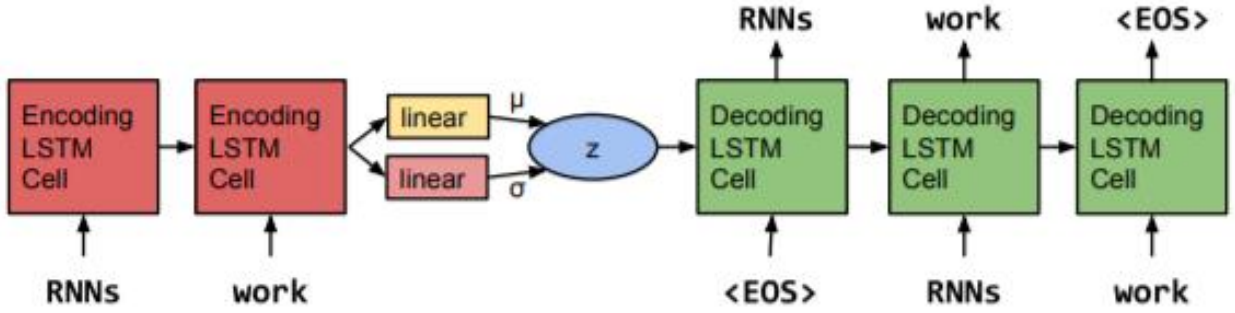


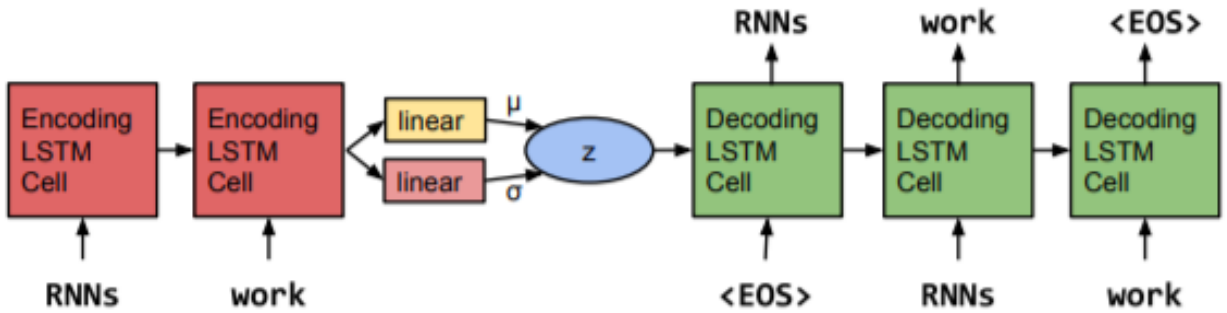
Figure 1: The core structure of our variational autoencoder language model. Words are represented using a learned randomly-initialized dictionary of embedding vectors. $\vec{z}$ is a vector-valued latent variable with a Gaussian prior and an approximate posterior parameterized by the encoder's outputs μ and σ. <EOS> marks the end of each sequence.

i went to the store to buy some groceries .
i store to buy some groceries .
i were to buy any groceries .
horses are to buy any groceries .
horses are to buy any animal .
horses the favorite any animal .
horses the favorite favorite animal .
horses are my favorite animal .

Table 1: Sentences produced by greedily decoding from points between two sentence encodings with a conventional autoencoder. The intermediate sentences are not plausible English.

$$\begin{aligned}\mathcal{L}(\theta; x) &= -\mathrm{KL}(q_\theta(\vec{z}|x)||p(\vec{z})) \\ &\quad + \mathbb{E}_{q_\theta(\vec{z}|x)}[\log p_\theta(x|\vec{z})] \\ &\leq \log p(x) \quad .\end{aligned}$$

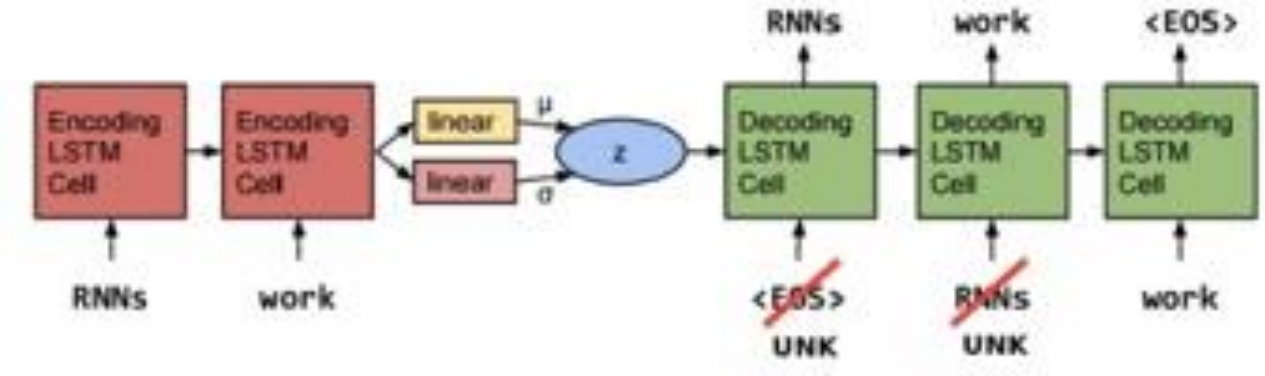
这篇文章实验中发现了问题：KL项很好优化，训练开始后，重构误差没怎么动，KL项很快就到0了。然后解码器性能又很强，再去降低重构误差，导致隐空间到头来还是没什么秩序



解决模型一开始不管重构误差，光优化KL的问题：KL cost annealing（KL成本退火）

给loss中的KL项加个权重，一开始为0，然后随训练轮数逐渐增加

解决解码器性能太强的问题：不给解码器input，让它去多多依赖隐空间的z

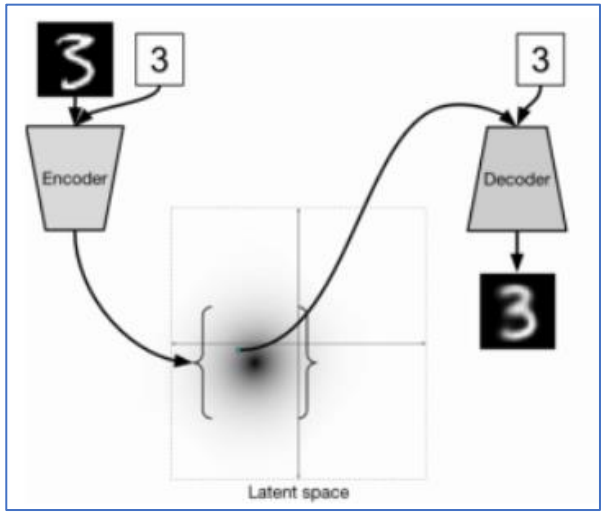


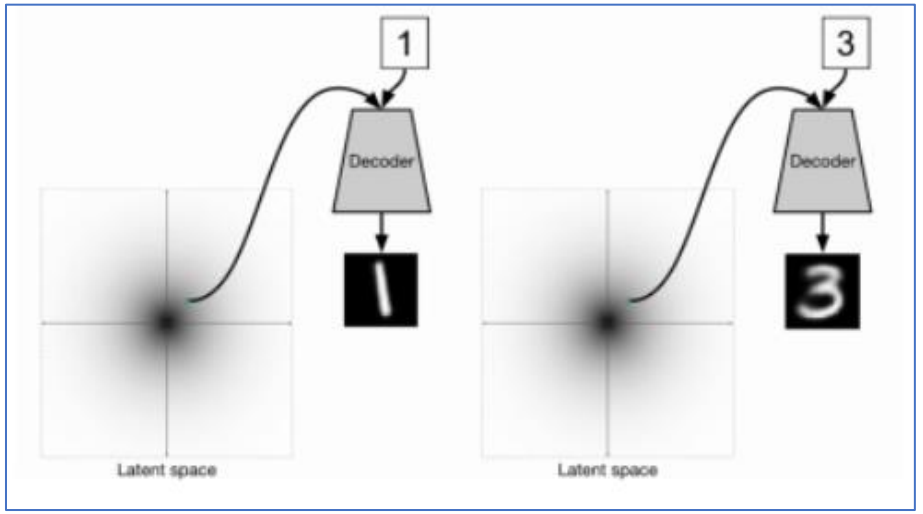
迷惑预警！

- 接下来会借助论文截图等介绍VAE的各种研究和应用
- 因为内容太多了所以内容比较粗糙
- 请大家把主要关注点放在“VAE都能做什么”上

VAE变种

条件VAE（Conditional VAE）





离散VAE

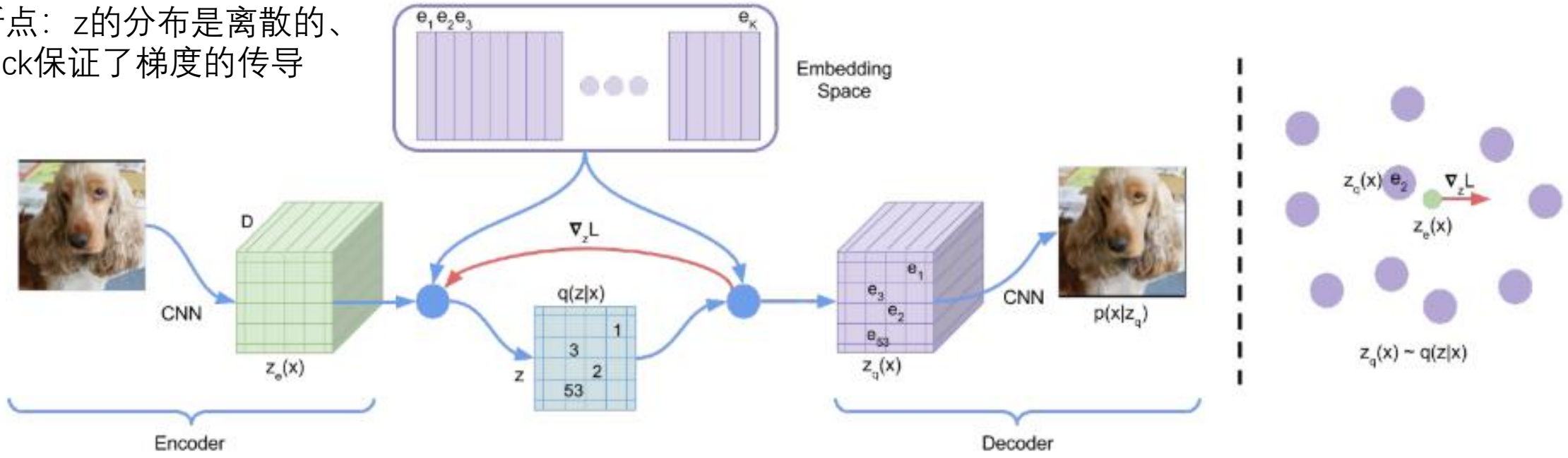
Neural Discrete Representation Learning

Aaron van den Oord
DeepMind
avdnoord@google.com

Oriol Vinyals
DeepMind
vinyals@google.com

Koray Kavukcuoglu
DeepMind
korayk@google.com

任务：和VAE一样
创新点：z的分布是离散的、用trick保证了梯度的传导



解纠缠VAE（Disentangled Variational Autoencoder）

解纠缠表示（隐空间表示的一种规则化形态）：意味着单个潜在单元对单个生成因素的变化敏感
vanilla VAE本身促使着q(z|x)上的后验分布更接近各向同性的标准高斯分布，所以本身有一定的解缠性

$$\mathcal{L} = \mathbb{E}_{q(z|X)}[\log p(X|z)] - D_{KL}[q(z|X)||p(z)]$$

然而上式的第一项的学习压力导致解缠性可能不太够。为了解决这个问题，有人给第二项加权，称之为β-VAE。但是这种方法如果给第二项加大权重会导致重构误差大，生成图片不精细。
为了使重构精度和解缠性质都好，大家就提出了新的KLD项，例如下面两个

$$D_{KL}(q(z|X)||p(z)) = I_q(z;n) + D_{KL}(q(z)||p(z))$$

$$D_{KL}(q(z)||p(z)) = D_{KL}(q(z)||\prod_j q(z_j)) + \sum_j D_{KL}(q(z_j)||p(z_j))$$

这些方法全部合体就是这样：$\mathcal{L} = \mathbb{E}_{q(z|X)}[\log p(X|z)] - I_q(z;n) - \beta D_{KL}(q(z)||\prod_j q(z_j)) - \sum_j D_{KL}(q(z_j)||p(z_j))$

https://towardsdatascience.com/disentanglement-with-variational-autoencoder-a-review-653a891b69bd

近年研究方向

近年研究方向

按VAE的形态分类

- 半监督学习
- 条件VAE
- Dual VAE
- 离散VAEVAE性能优化
- Few-shot学习
- 分级(hierarchical)VAE

按VAE的功能分类

- 一般生成
- 成对图像生成
- 成对文本生成
  - QA
  - if-then
- 上文回复生成
- 长文本生成
- 表示增强
- 图信息编码

表示増強

Variational Pretraining for Semi-supervised Text Classification

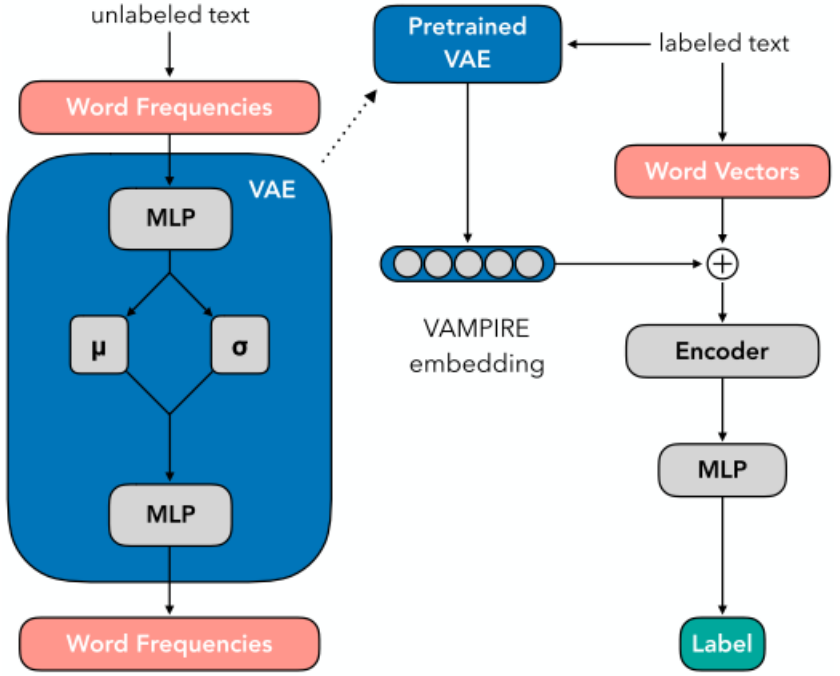
ACL2020

Suchin Gururangan[1] **Tam Dang**[2] **Dallas Card**[3] **Noah A. Smith**[1,2]

[1]Allen Institute for Artificial Intelligence, Seattle, WA, USA
[2]Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, WA, USA
[3]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA, USA
suching@allenai.org {dangt7,nasmith}@cs.washington.edu dcard@cmu.edu

任务：文本分类
预训练输入：无标记文本
预训练输出：无标记文本
分类器输入：预训练模型隐层输出(VAMPIRE embedding)+文本向量
分类器输出：标签

VAE学到的是一些辅助信息



Do sequence-to-sequence VAEs learn global features of sentences?

EMNLP2020

Tom Bosc
Mila, Université de Montréal
bosct@mila.quebec

Pascal Vincent
Mila, Université de Montréal, CIFAR
vincentp@iro.umontreal.ca

探讨了用VAE编码句子能否更好的学习句子的全局信息。例如：情感、主题。

实验方法：分解句子中每个部分的重构损失。

实验结果：VAE容易记住句子的第一个单词和句子长度。

其他贡献：研究了基于词袋假设和语言模型预训练的代替体系结构，这些变体可以学习到更具全局性的潜在变量，即更具预测性的话题或情感标签。

VCDM: Leveraging Variational Bi-encoding and Deep Contextualized Word Representations for Improved Definition Modeling

EMNLP2020

Machel Reid **Edison Marrese-Taylor** **Yutaka Matsuo**
Graduate School of Engineering
The University of Tokyo
machelreid2004@gmail.com
{emarrese, matsuo}@weblab.t.u-tokyo.ac.jp

任务：单词表示、定义生成
输入一个单词，输出词典风格的定义。

训练集：牛津大辞典、Wikipedia、剑桥词典等词典

效果如图：

Word	Frankenstein
Context	In arming the dictator, the US was creating a ***Frankenstein***
Reference	something that destroys or harms the person or people who created it

Generated	BL	P	R	F
something that you say or do that you think someone of something is ridiculous	**12.5**	83.41	84.78	84.09
an extremely frightening or offensive person	8.13	**87.00**	84.78	**85.88**

EMNLP2020

Public Sentiment Drift Analysis Based on Hierarchical Variational Auto-encoder

Wenyue Zhang[1], Xiaoli Li[2], Yang Li[1], Suge Wang[1], Deyu Li[1], Liao Jian[1], Jianxing Zheng[1]
[1]School of Computer and Information Technology, Shanxi University, Taiyuan 030006, China
[2]Institute for Infocomm Research, Singapore
wsg@sxu.edu.cn

任务：public sentiment drift analysis
输入：大量的公众文章，按时间排序
输出：不同时间公众情感的分布，包括预测未来的公众情感分布

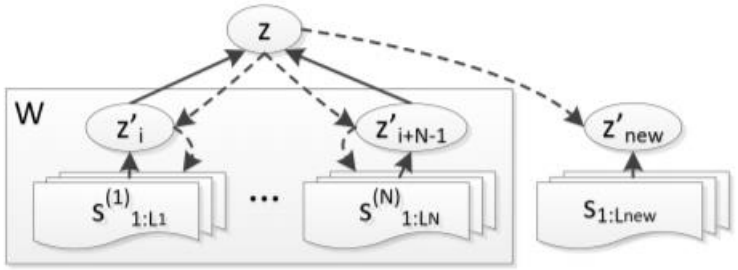


Table 1: Annotations of our method

W	N length slide window.
s	A sentiment document, with its superscript indicating its time period.
L_i	Data quantity in i'th period.
z', z	Latent meta-distributions of period and window, respectively.
θ, ϕ	Parameters of decoder and encoder.
dash/solid line	Decode/Encode process.

方法：按时间顺序对文章进行编码学习，学习后再按照时间顺序依次生成编码分布，然后对比分析不同时间分布差异即可

$$(\mu'_{\phi_i}, \sigma'_{\phi_i}) = \text{EncoderModel}_\phi(s^{(i)}_{1:L_i}) \quad (1)$$

$$z'|s^{(i)}_{1:L_i} \sim N(\mu'_{\Phi_i}, \sigma'^2_{\Phi_i}) \quad (2)$$

$$(\mu_\Phi, \sigma_\phi) = \text{EncoderModel}_\phi(z'_{1:N}) \quad (3)$$

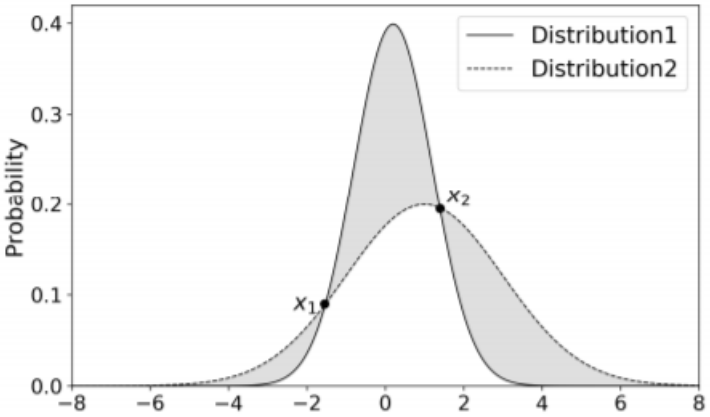
$$z|z'_{1:N} \sim N(\mu_\phi, \sigma^2_\phi) \quad (4)$$

$$(\mu_\theta, \sigma_\theta) = \text{DecoderModel}_\theta(z) \quad (5)$$

$$z'_i|z \sim N(\mu_\theta, \sigma^2_\theta) \quad (6)$$

$$(\mu'_{\theta_i}, \sigma'_{\theta_i}) = \text{DecoderModel}_\theta(z'_i) \quad (7)$$

$$s^{(i)}_j|z'_i \sim N(\mu'_{\theta_i}, \sigma'^2_{\theta_i}) \quad (8)$$



直观理解分布差异（阴影部分）

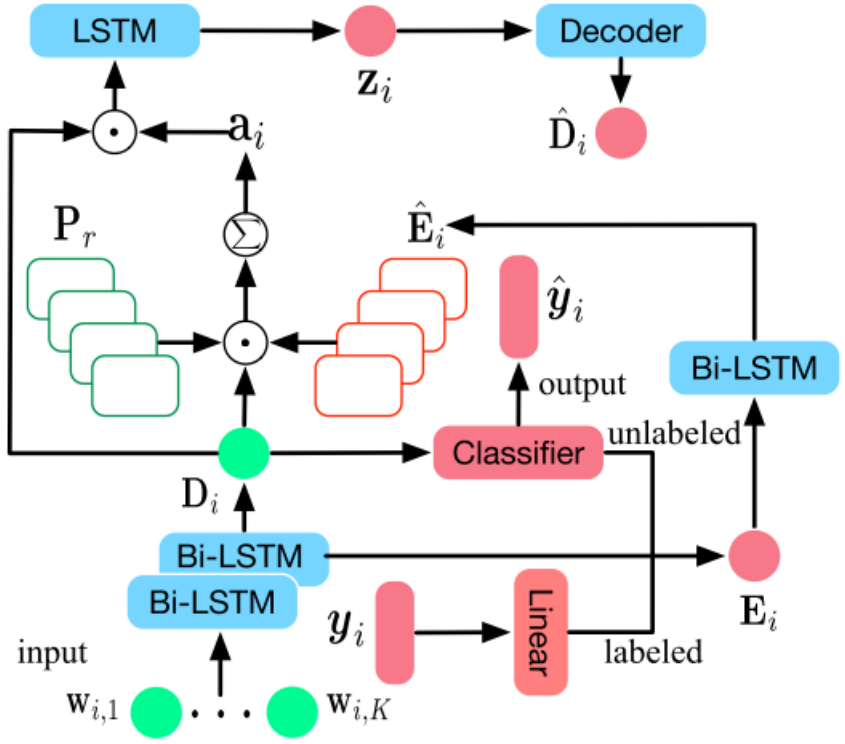
$$\log p(S) \geq E_{q_\phi(z'_{1:N}, z|S)}\Big[\overbrace{\log p(z) + \log p_\theta(z'_{1:N}|z) + \log p_\theta(S|z'_{1:N})}^{Decode} \underbrace{-\log q_\phi(z|z'_{1:N}) - \log q_\phi(z'_{1:N}|S)}_{Encode}\Big] \quad (9)$$

半监督学习

Interpretable Operational Risk Classification with Semi-Supervised Variational Autoencoder

ACL2020

Fan Zhou[1], **Shengming Zhang**[1], **Yi Yang**[2]
[1]University of Electronic Science and Technology of China.
[2]Hong Kong University of Science and Technology.
fan.zhou@uestc.edu.cn, shmizhang@gmail.com, imyiyang@ust.hk



VAE搭配multi-head attention做半监督的Operational Risk Classification金融领域
任务：多标签分类
输入：一篇文章
输出：是否与数据隐私有关、是否与银行起诉有关、是否与xxx有关等等，所有答案用一整个向量表示，每一维跟一类标签相关，大小为0~1

NIPS2020

任务：二分类
使用数据：只有正例和未标记数据
创新点：不受VAE先验分布的限制，可以直接从给定数据中定量地评估贝叶斯分类器的建模误差

A Variational Approach for Learning from Positive and Unlabeled Data

Hui Chen[*]
School of Mathematical Science
Tongji University, Shanghai, P. R. China
hui.chen96@outlook.com

Fangqing Liu[*]
School of Mathematical Science
Tongji University, Shanghai, P. R. China
fangqingliu0@gmail.com

Yin Wang
School of Electronics and Information Engineering
Tongji University, Shanghai, P. R. China
yinw@tongji.edu.cn

Liyue Zhao
Cloudwalk Inc.
Shanghai, P. R. China
zhaoliyue@cloudwalk.cn

Hao Wu[†]
School of Mathematical Science
Tongji University, Shanghai, P. R. China
hwu@tongji.edu.cn

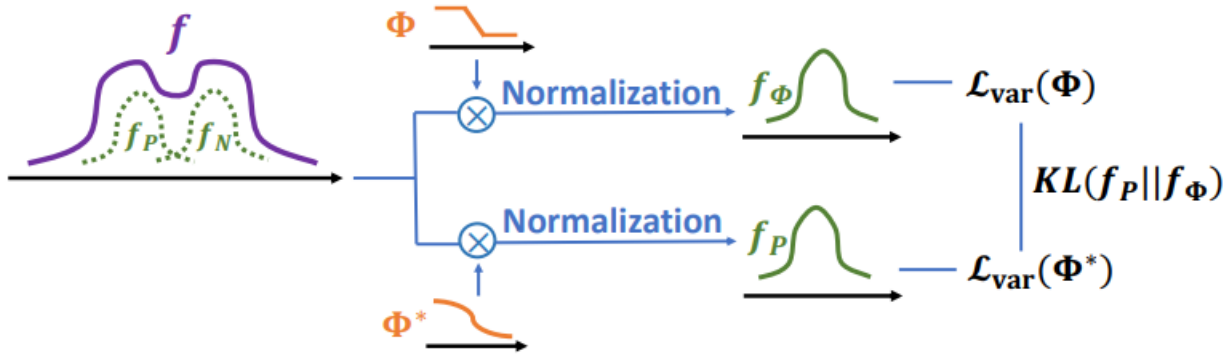


Figure 1: Graphical interpretation of the variational principle stated by Theorem 3, where f_P, f_N, f denote distributions of positive, negative and unlabeled data. Each classifier model Φ induces an approximation f_Φ of f_P as in (4), and $\mathrm{KL}(f_P||f_\Phi)$ equals to the difference between functionals $\mathcal{L}_{\text{var}}(\Phi)$ and $\mathcal{L}_{\text{var}}(\Phi^*)$.

图像生成

NVAE: A Deep Hierarchical Variational Autoencoder

NIPS2020

Arash Vahdat, Jan Kautz
NVIDIA
{avahdat, jkautz}@nvidia.com

利用各种卷积、归一化、残差、谱正则之类的花里胡哨提高了VAE的生成图片，带头在256*256这种超大图片生成任务上做出了很好的效果

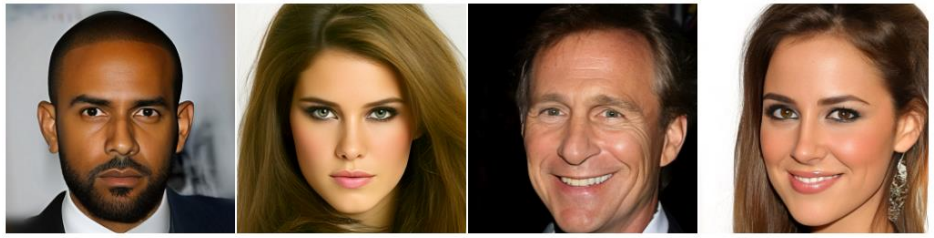
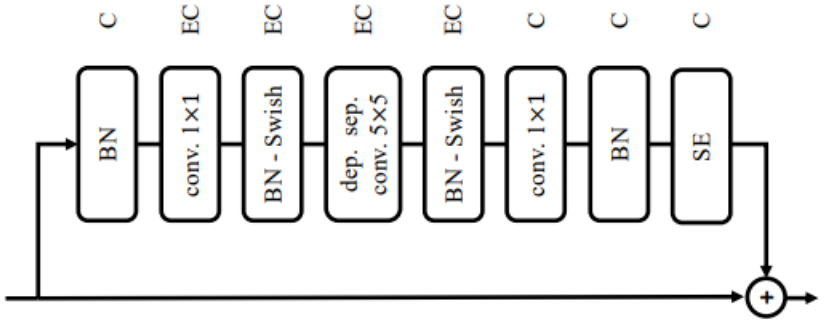
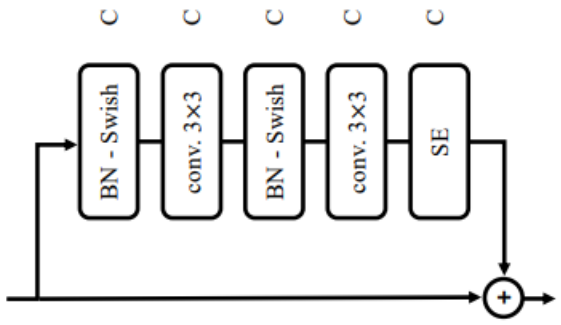


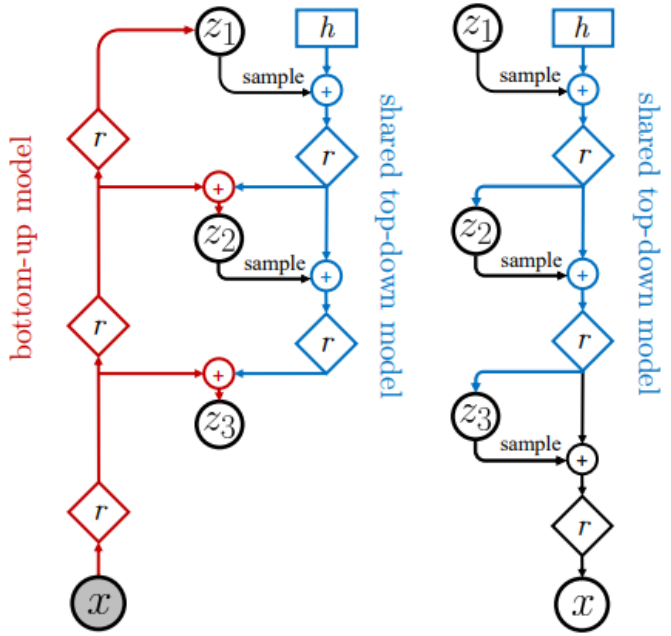
Figure 1: 256×256-pixel samples generated by NVAE, trained on CelebA HQ [28]



(a) Residual Cell for NVAE Generative Model



(b) Residual Cell for NVAE Encoder



(a) Bidirectional Encoder (b) Generative Model

Figure 2: The neural networks implementing an encoder $q(\boldsymbol{z}|\boldsymbol{x})$ and generative model $p(\boldsymbol{x}, \boldsymbol{z})$ for a 3-group hierarchical VAE. ◇r denotes residual neural networks, ⊕ denotes feature combination (e.g., concatenation), and h is a trainable parameter.

Dual VAE

Crossing Variational Autoencoders for Answer Retrieval

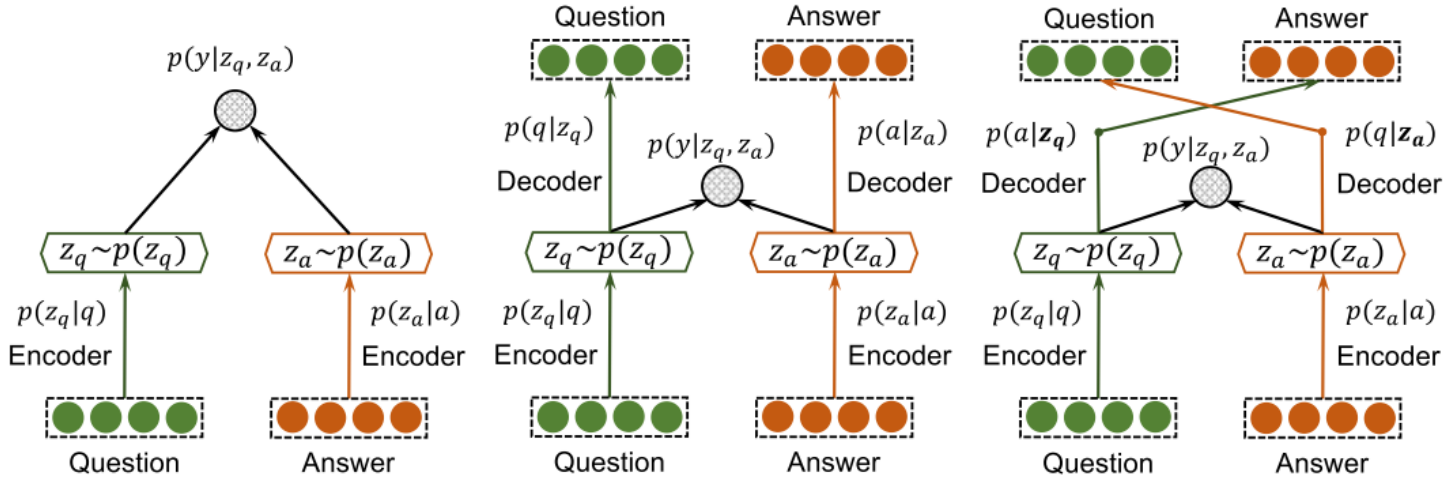
ACL2020

Wenhao Yu†, Lingfei Wu‡, Qingkai Zeng†, Shu Tao‡, Yu Deng‡, Meng Jiang†
†University of Notre Dame, Notre Dame, IN, USA
‡IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA
†{wyu1, qzeng, mjiang2}@nd.edu
‡{wuli, shutao, dengy}@us.ibm.com

任务：Answer Retrieval。输入：问句和一组答句。输出：正确的答句。
文章特色：让答句和问句对齐。

Question (1): What three stadiums did the NFL decide between for the game?
Question (2): What three cities did the NFL consider for the game of Super Bowl 50?
...
Question (17): How many sites did the NFL narrow down Super Bowl 50's location to?

Answer: The league eventually narrowed the bids to three sites: New Orleans Mercedes-Benz Superdome, Miami Sun Life Stadium, and the San Francisco Bay Area's Levi's Stadium.



(a) Dual-Encoders (Yang et al., 2019) (b) Dual-VAEs (Shen et al., 2018) (c) Dual-CrossVAEs (**Ours**)

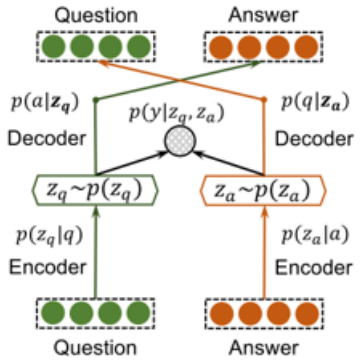
Figure 1: (a)–(b) The Q-A alignment and Q/A semantics were learned too separately to capture the aligned semantics between question and answer. (c) We propose to cross VAEs by generating questions with aligned answers and generating answers with aligned questions.

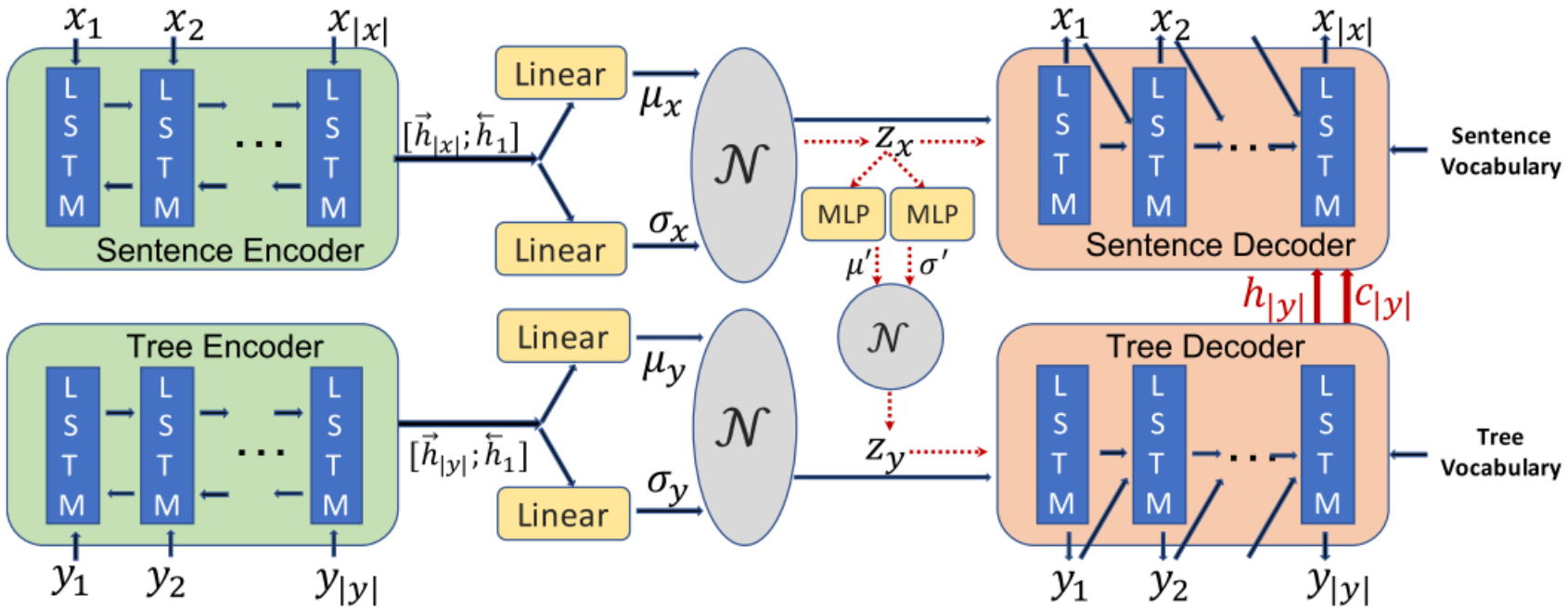
Syntax-Infused Variational Autoencoder for Text Generation

ACL2019

Xinyuan Zhang[1]*, Yi Yang[2]*, Siyang Yuan[1], Dinghan Shen[1], Lawrence Carin[1]
[1]Duke University
[2]ASAPP Inc.
xy.zhang@duke.edu, yyang@asapp.com

任务：文本生成
创新点：分别对句子和句子的语法树进行操作，生成时，可给定语法树，生成对应该语法树的语句。





NIPS2019

任务：生成同一张人脸的异构图像，比如抬头和低头、黑白和彩色、红外图像和偏振热图像等，用于异构人脸识别的训练。现存的模型通常是从一张图片生成另一张图片，类似于翻译，生成的图片数量有限，且多样性差，一致性也可能较差

异构人脸识别：数据库中有一个人的黑白照片，他拍了张红外图像我照样能认出他是谁

动机：异构人脸识别训练数据太少

贡献：每次生成同一个人的异构的一对照片

Dual Variational Generation for Low Shot Heterogeneous Face Recognition

Chaoyou Fu[1,2]*, **Xiang Wu**[1]*, **Yibo Hu**[1], **Huaibo Huang**[1], **Ran He**[1,2,3]†
[1]NLPR & CRIPAC, CASIA
[2]University of Chinese Academy of Sciences
[3]Center for Excellence in Brain Science and Intelligence Technology, CAS
{chaoyou.fu, rhe}@nlpr.ia.ac.cn, alfredxiangwu@gmail.com
{yibo.hu, huaibo.huang}@cripac.ia.ac.cn

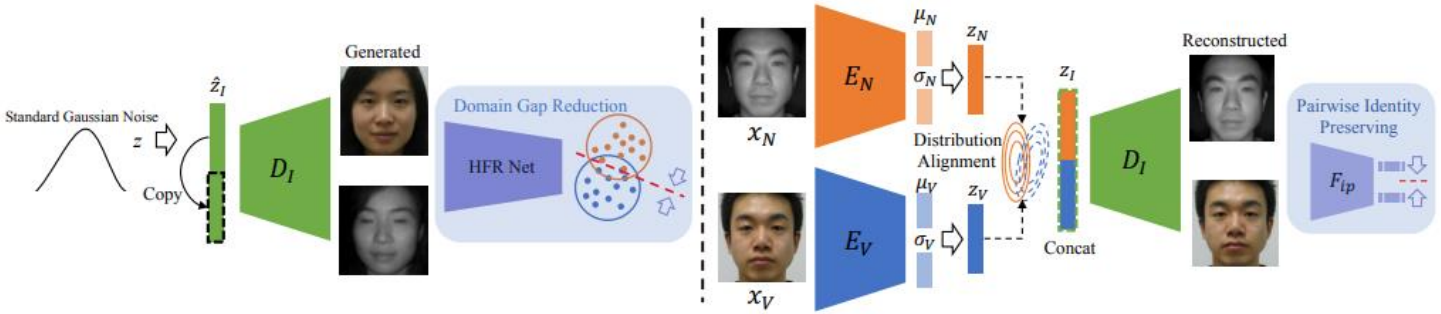


Figure 3: The purpose (left part) and training model (right part) of our unconditional DVG framework. DVG generates large-scale new paired heterogeneous images with the same identity from standard Gaussian noise, aiming at reducing the domain discrepancy for HFR. In order to achieve this purpose, we elaborately design a dual variational autoencoder. Given a pair of heterogeneous images from the same identity, the dual variational autoencoder learns a joint distribution in the latent space. In order to guarantee the identity consistency of the generated paired images, we impose a distribution alignment in the latent space and a pairwise identity preserving in the image space.

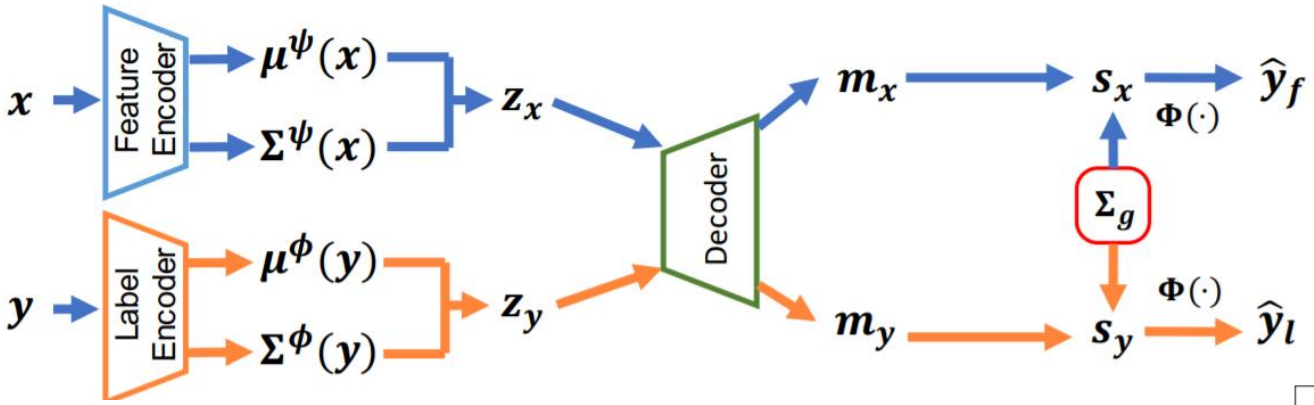
Disentangled Variational Autoencoder based Multi-Label Classification with Covariance-Aware Multivariate Probit Model

AAAI2021

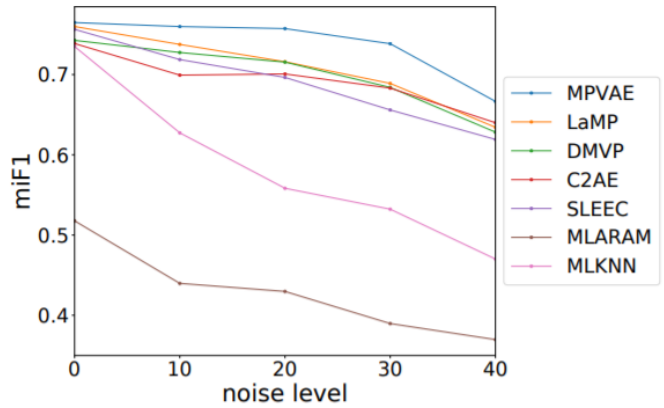
Junwen Bai*, Shufeng Kong and Carla Gomes
Department of Computer Science, Cornell University
{jb2467, sk2299}@cornell.edu, gomes@cs.cornell.edu

任务：多标签分类

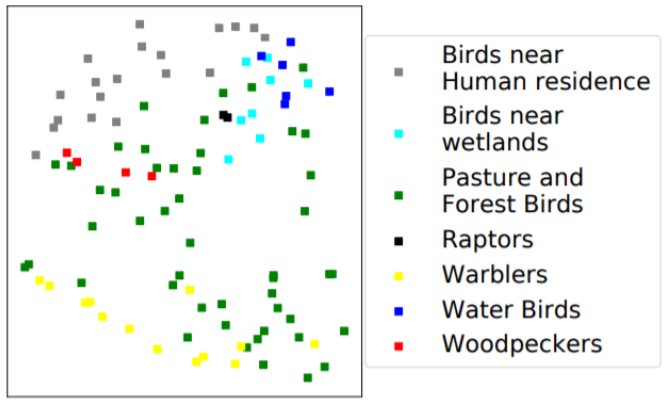
部分摘要：多元概率位变分自动编码器（MPVAE）能有效地学习潜在嵌入空间和标签相关性。MPVAE学习并对齐标签和特征的两个概率嵌入空间。MPVAE解码器从嵌入空间通过学习共享协方差矩阵，在多变量概率模型下对输出目标的联合分布进行建模。该模型的一个特点是抗噪声能力强（例如错误标注的样本）



注：zx和zy被设定为维数相同，它们共用一个解码器，Σg是一个全局的协方差矩阵，sx、sy是从N(mx,Σg)、N(my,Σg)中采样得到的。在测试过程中，只是用ŷf作为输出



抗干扰效果示意图



聚类效果示意图

数据集：
eBird
与鸟和鸟可能的栖息地相关

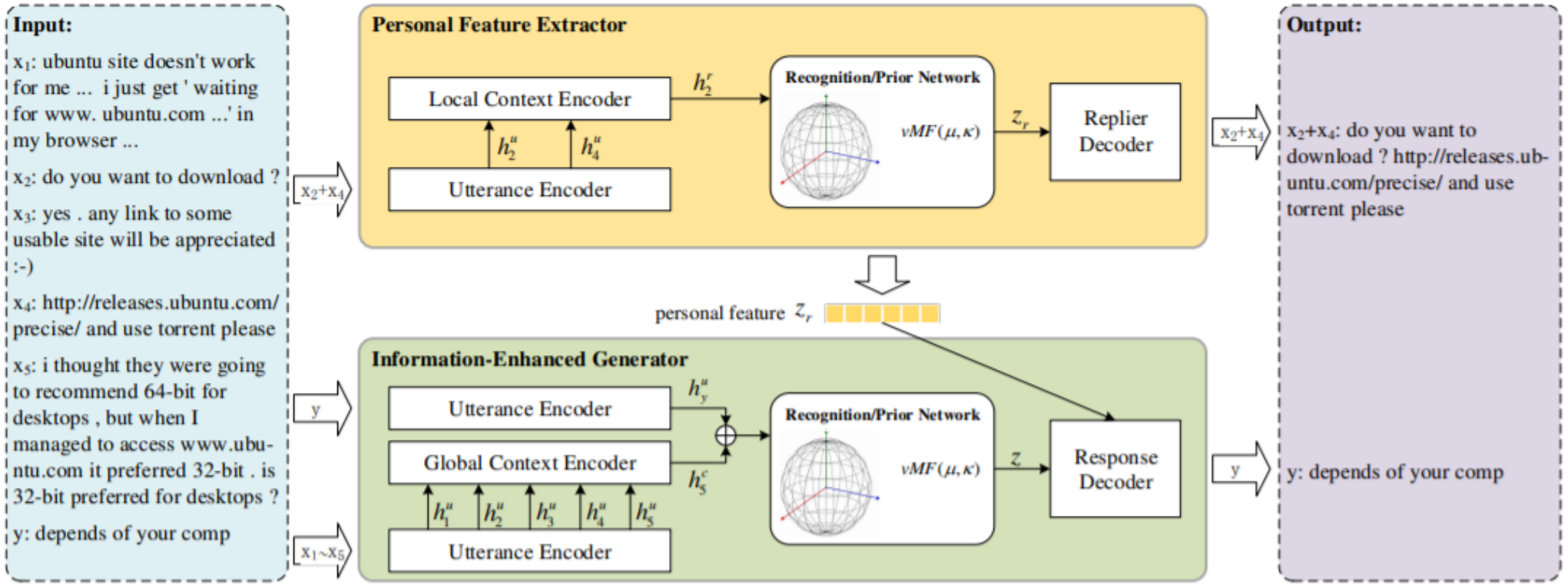
EMNLP2019

任务：对话生成
动机：1.latent space futility. 2.replier-consistency decay
注：图中Input是双人对话
创新点：1.用了vMF分布代替高斯。2.抽取回复者说话风格用于生成回答

A Semi-Supervised Stable Variational Network for Promoting Replier-Consistency in Dialogue Generation

Jinxin Chang[1,2,*], Ruifang He[1,2,3,*†], Longbiao Wang[1,2,†], Xiangyu Zhao[1,2], Ting Yang[1,2], and Ruifang Wang[1,2]

[1]Tianjin Key Laboratory of Cognitive Computing and Application, Tianjin, China.
[2]College of Intelligence and Computing, Tianjin University, Tianjin, China.
[3]State Key Laboratory of Cognitive Intelligence, iFLYTEK, P.R. China.
{changjinxin,rfhe,longbiao_wang}@tju.edu.cn
{zhaoxiangyu009,16622898776,wrf276224255}@163.com



任务详述：A问一句，B答一句，A又问，B又答，A又问，B又答，A又问，请生成B的回答。

黄色部分会把B已知的所有回答都放进去。

Idea

- dual VAE可以捕获现实中的对偶关系
- 那我们可以做个triple VAE，捕获现实中的三元关系
- 最容易想到的就是EToDs里面的数据库了，都是三元组
- 这样就可以用VAE对EToDs训练集的数据库做生成增广了

条件VAE（CVAE）

ACL2020

Generating Diverse and Consistent QA pairs from Contexts with Information-Maximizing Hierarchical Conditional VAEs

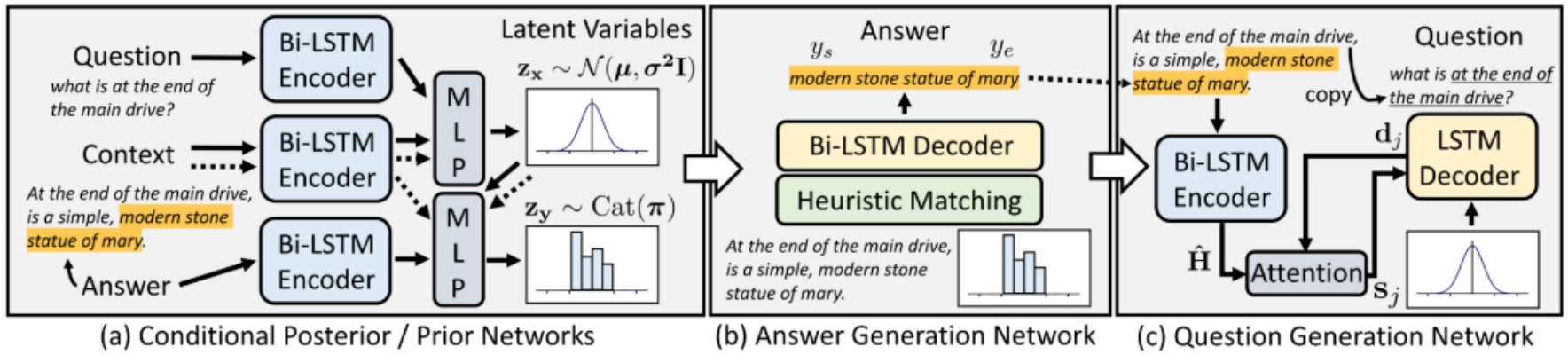
Dong Bok Lee[1*] **Seanie Lee**[1,3*] **Woo Tae Jeong**[3] **Donghwan Kim**[3] **Sung Ju Hwang**[1,2]
KAIST[1], AITRICS[2], 42Maru Inc.[3], South Korea
{markhi,lsnfamily02,sjhwang82}@kaist.ac.kr
{wtjeong,scissors}@42maru.com

输入：一段话
输出：一个QA对
动机：解决QA系统训练数据少的问题

Paragraph (Input) Philadelphia has more murals than any other u.s. city, thanks in part to the 1984 creation of the department of recreation's mural arts program, ...The program has funded more than 2,800 murals
Q1 which city has more murals than any other city? **A1** philadelphia
Q2 why philadelphia has more murals? **A2** the 1984 creation of the department of recreation's mural arts program
Q3 when did the department of recreation' s mural arts program start ? **A3** 1984
Q4 how many murals funded the graffiti arts program by the department of recreation? **A4** more than 2,800

Table 1: An example of QA pairs generated with our framework. The paragraph is an extract from Wikipedia provided by Du and Cardie (2018). For more examples, please see Appendix D.

$$p_\theta(\mathbf{x},\mathbf{y}|\mathbf{c}) = \int_{\mathbf{z_x}} \sum_{\mathbf{z_y}} p_\theta(\mathbf{x}|\mathbf{z_x},\mathbf{y},\mathbf{c}) p_\theta(\mathbf{y}|\,\mathbf{z_x},\mathbf{z_y},\mathbf{c}) \cdot p_\psi(\mathbf{z_y}|\mathbf{z_x},\mathbf{c}) p_\psi(\mathbf{z_x}|\mathbf{c}) d\mathbf{z_x}$$



Pre-train and Plug-in: Flexible Conditional Text Generation with Variational Auto-Encoders

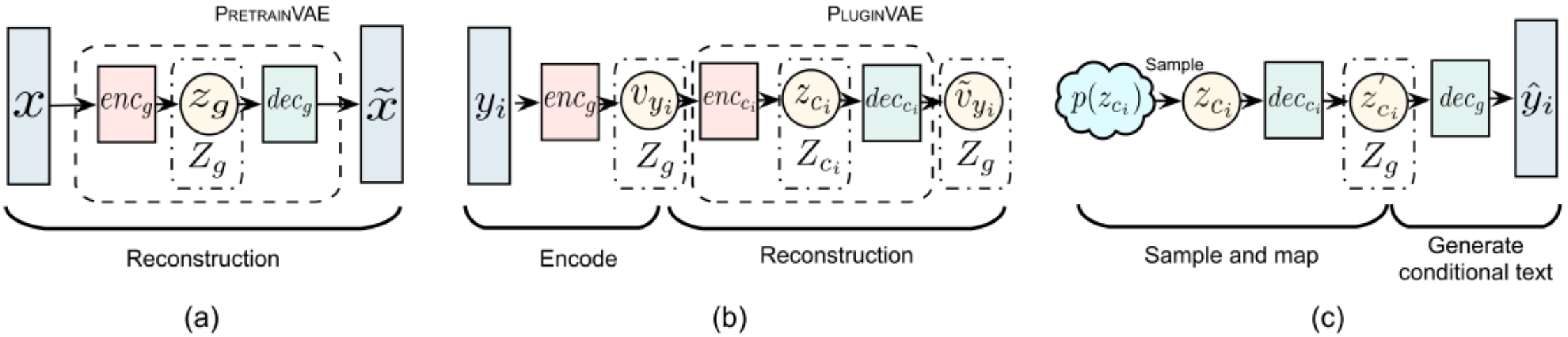
ACL2019

Yu Duan[1*], **Canwen Xu**[2*], **Jiaxin Pei**[3*], **Jialong Han**[4†], **Chenliang Li**[2‡]
[1] Alibaba Group, China [2] Wuhan University, China
[3] University of Michigan, United States [4] Amazon, United States
[1] derrick.dy@alibaba-inc.com, [2] {xucanwen,cllee}@whu.edu.cn
[3] pedropei@umich.edu, [4] jialonghan@gmail.com

任务：使用VAE的条件文本生成。
创新点：基于端到端的文本生成在不同情况下需要完全重新训练。
本文模型在遇到新情况(new condition)时，只需重新训练网络上的一个小插件，然后插在整个模型上就行了。

模型分为两部分。（1）Pre-train VAE，用来general地训练一个VAE。（2）Plug-in VAE，训练目标是实现条件隐空间（conditional latent space）和全局隐空间（global latent space）的相互映射。

~~setting in (Hu et al., 2017).~~ Given a set of k condi-
tions C
and con
where e
the con
learn a
able z
bution
conditi
able z
realistic text samples matching the given condition.

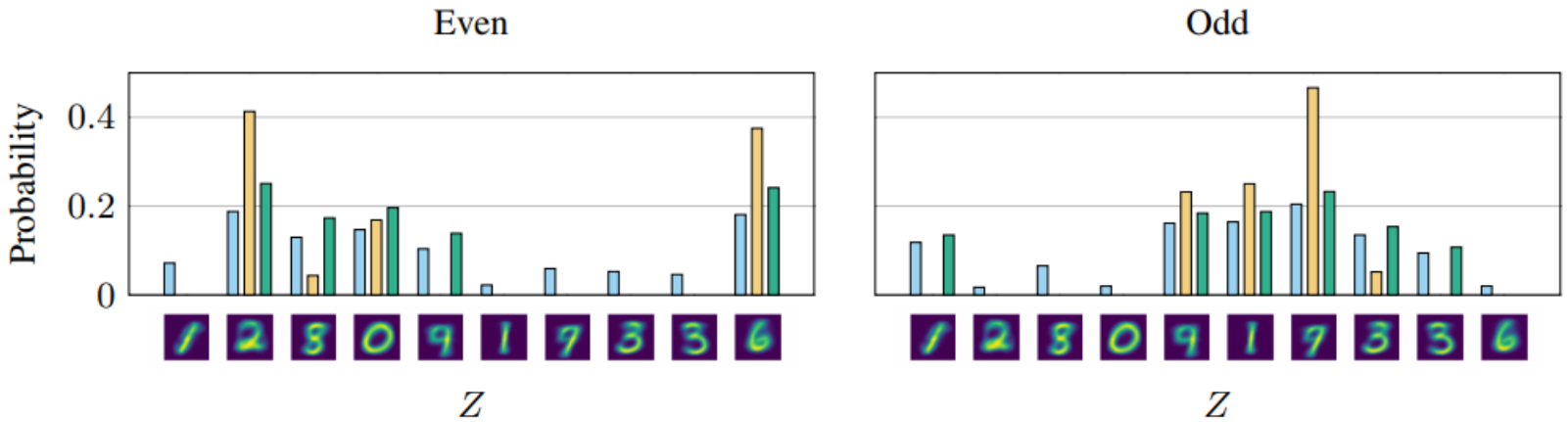


Evidential Sparsification of Multimodal Latent Spaces in Conditional Variational Autoencoders

NIPS2020

Masha Itkina, Boris Ivanovic, Ransalu Senanayake, Mykel J. Kochenderfer, Marco Pavone
Department of Aeronautics and Astronautics
Stanford University
{mitkina, borisi, ransalu, mykel, pavone}@stanford.edu

离散VAE，隐空间为10类，带着奇偶信息训练，训练的时候走上面，测试的时候走下面，最后结果如下图



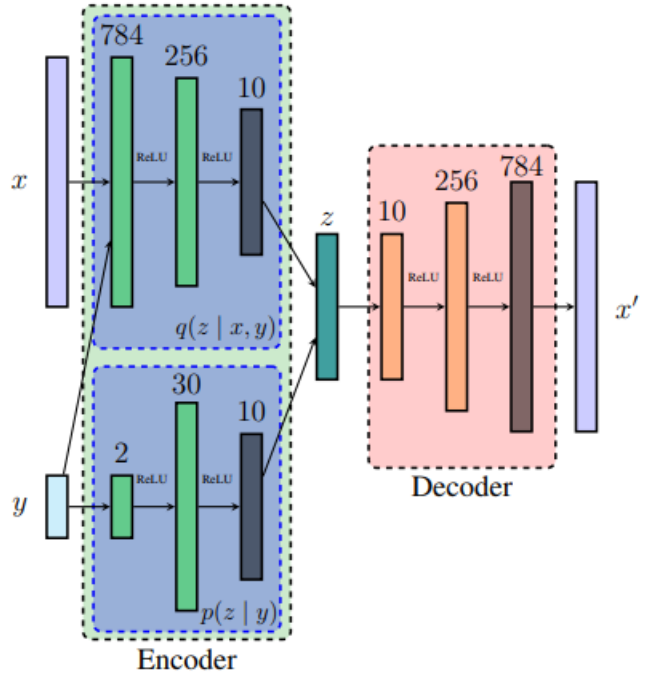


Figure 1: The CVAE architecture used for MNIST image generation. The last layer in each MLP is a softmax layer. At test time, $p(z \mid y)$ is used to sample the latent space.

EMNLP2019

A Discrete CVAE for Response Generation on Short-Text Conversation

Jun Gao[1]*, Wei Bi[2]†, Xiaojiang Liu[2], Junhui Li[1], Guodong Zhou[1], Shuming Shi[2]
[1]School of Computer Science and Technology, Soochow University, Suzhou, China
imgaojun@gmail.com, {lijunhui,gdzhou}@suda.edu.cn
[2]Tencent AI Lab, Shenzhen, China
{victoriabi, kieranliu, shumingshi}@tencent.com

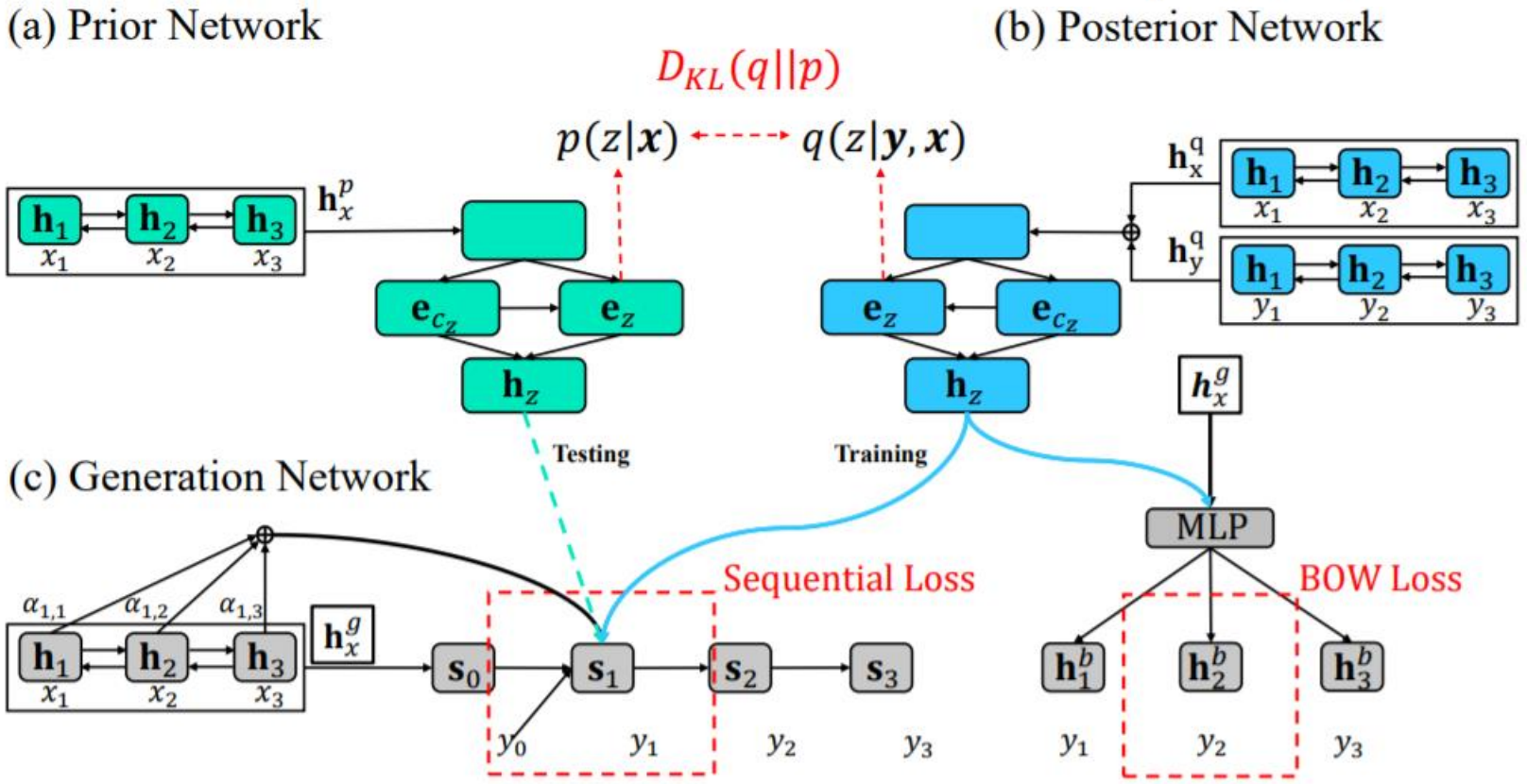


Figure 1: The architecture of the proposed discrete CVAE. $\mathbf{e}_{c_z}$ and $\mathbf{e}_z$ are embeddings of a cluster and a word sampled from the estimated discrete distributions. $\mathbf{e}_{c_z}$ is only applied when the two-stage sampling approach in Section 3.2 is used. If $\mathbf{e}_{c_z}$ is applied, the latent representation $\mathbf{h}_z$ is the sum of $\mathbf{e}_{c_z}$ and $\mathbf{e}_z$; otherwise, $\mathbf{h}_z$ is $\mathbf{e}_z$. α denotes the attention weight. $\oplus$ denotes the sum of input vectors.

任务：回复生成
输入：一句话
输出：这句话的答复

训练：x，y共同建模z的分布（称为分布1），然后用x编码z（称为分布2），要求分布2模拟分布1

离散VAE

NIPS2019

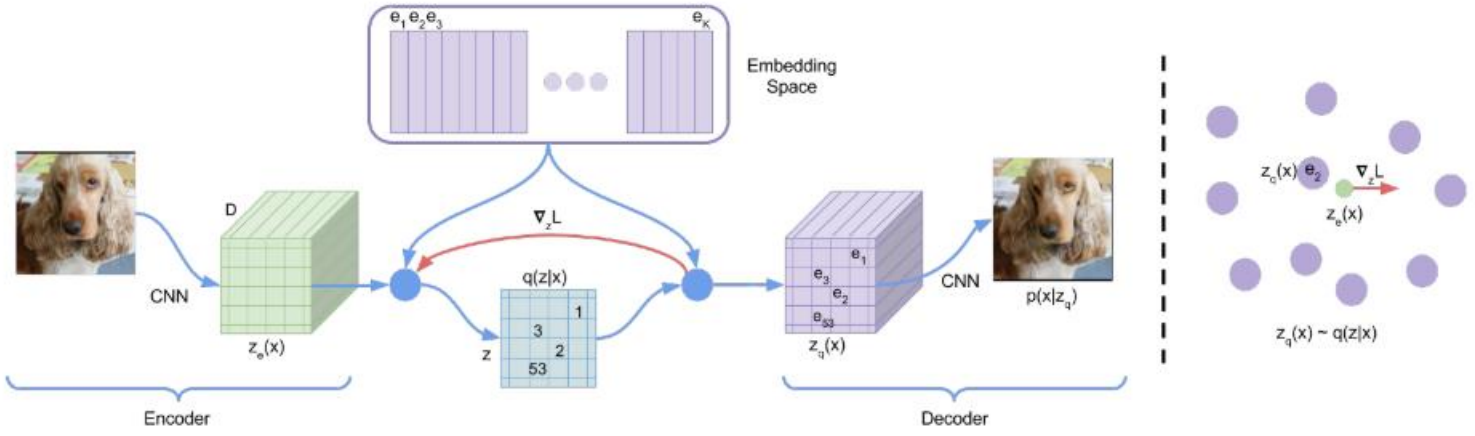
Direct Optimization through arg max for Discrete Variational Auto-Encoder

Guy Lorberbom Technion
Andreea Gane MIT
Tommi Jaakkola MIT
Tamir Hazan Technion

In this work, we use the Gumbel-Max trick to reparameterize discrete VAEs using the arg max prediction and show how to propagate gradients through the non-differentiable arg max function

翻译：用另一种方法让离散隐空间VAE的梯度传回去

回想一下这个2017年的：



ACL2020

Evidence-Aware Inferential Text Generation with Vector Quantised Variational AutoEncoder

Daya Guo[1]*, Duyu Tang[2], Nan Duan[2], Jian Yin[1], Daxin Jiang[3] and Ming Zhou[2]
[1] The School of Data and Computer Science, Sun Yat-sen University.
Guangdong Key Laboratory of Big Data Analysis and Processing, Guangzhou, P.R.China
[2] Microsoft Research Asia, Beijing, China
[3] Microsoft Search Technology Center Asia, Beijing, China
{guody5@mail2,issjyin@mail}.sysu.edu.cn
{dutang,nanduan,djiang,mingzhou}@microsoft.com

任务：Inferential Text Generation，给定Event、Background，生成Inferences

创新点：利用VAE产生的隐变量去和Evidence(Background)算相似度，充分利用信息。

使用离散VAE。隐空间里就几个向量，VAE算出的是个像softmax似的东西，然后取最大值对应的那一条向量。

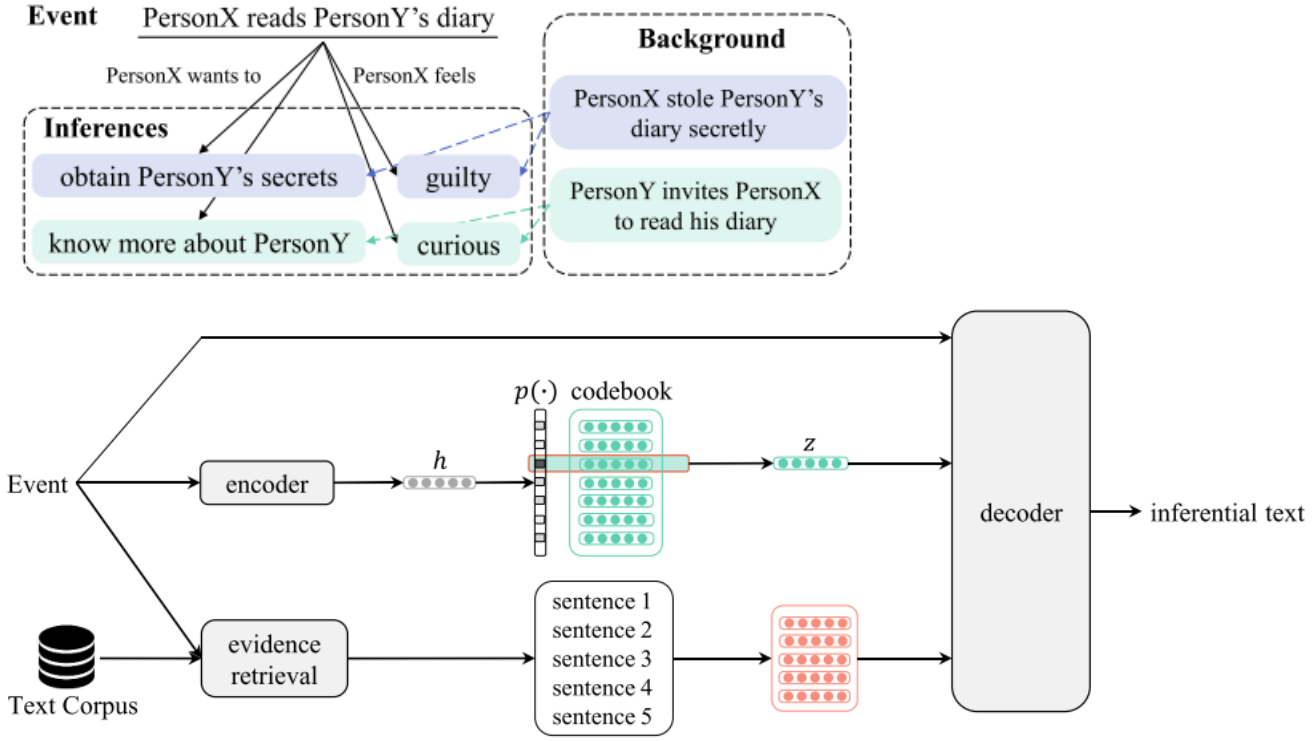


Figure 3: The model architecture of our approach.

EMNLP2019

这篇跟离散VAE无关，放在这跟上一篇比较一下
这一篇是给定event、inference，生成background

Modeling Event Background for If-Then Commonsense Reasoning Using Context-aware Variational Autoencoder

Li Du, Xiao Ding, Ting Liu* and Zhongyang Li
Research Center for Social Computing and Information Retrieval
Harbin Institute of Technology, China
{ldu, xding, tliu, zyli}@ir.hit.edu.cn

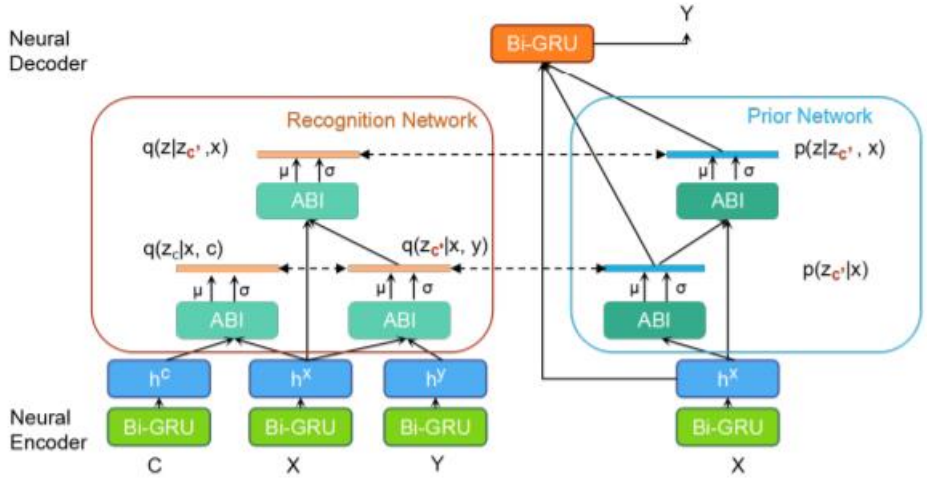


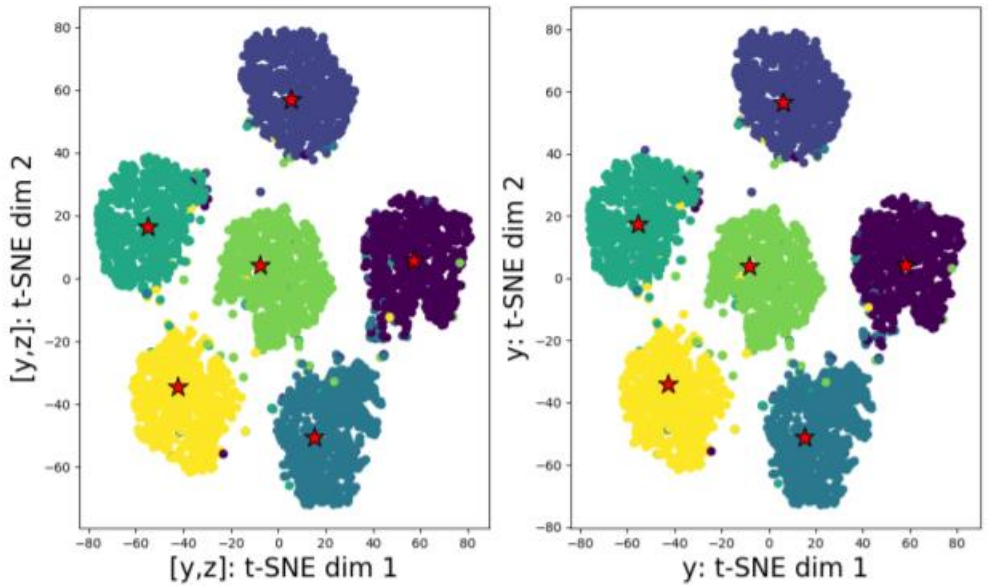
Figure 4: Architecture of CWVAE. We mark Neural encoder in green, prior network in blue, recognition network in brown and neural decoder in orange, respectively.

Base event	Inference Dim.	Target
PersonX adopts a child	xIntent	to help another person, to have a child
	xNeed	to visit adoption agency, to be approved for adoption
	xAttr	compassionate, generous
	xEffect	becomes a parent, gains love and companionship
	xWant	take child home, buy child clothes
	xReact	happy, caring
	oReact	has a parent, receives love and affection
	oWant	try on new clothes, to have a family
	oEffect	has a parent, Receives love and affection

说实话VAE的模型基本就是数据在模型里面的流动，然后加上复杂的要死的loss，让人好像看懂又好像看不懂。。
所以说上面的模型我反正没看懂哈哈

AAAI2021

任务：开集识别
输入：图片和一组类别
输出：图片所属的类别或图片不属于任何类别
使用：高斯混合VAE



编码视觉效果

Open-Set Recognition with Gaussian Mixture Variational Autoencoders

Alexander Cao
Department of Industrial Engineering and Management Sciences
Northwestern University
Evanston, IL 60208
a-cao@u.northwestern.edu

Yuan Luo
Department of Preventive Medicine
Northwestern University
Chicago, IL 60611
yuan.luo@northwestern.edu

Diego Klabjan
Department of Industrial Engineering and Management Sciences
Northwestern University
Evanston, IL 60208
d-klabjan@northwestern.edu

图信息编码

NIPS2019

VAE的一个分支：
把有向无环图信息
用设计过的VAE编
码到隐空间上

D-VAE: A Variational Autoencoder for Directed Acyclic Graphs

Muhan Zhang, Shali Jiang, Zhicheng Cui, Roman Garnett, Yixin Chen
Department of Computer Science and Engineering
Washington University in St. Louis
{muhan, jiang.s, z.cui, garnett}@wustl.edu, chen@cse.wustl.edu

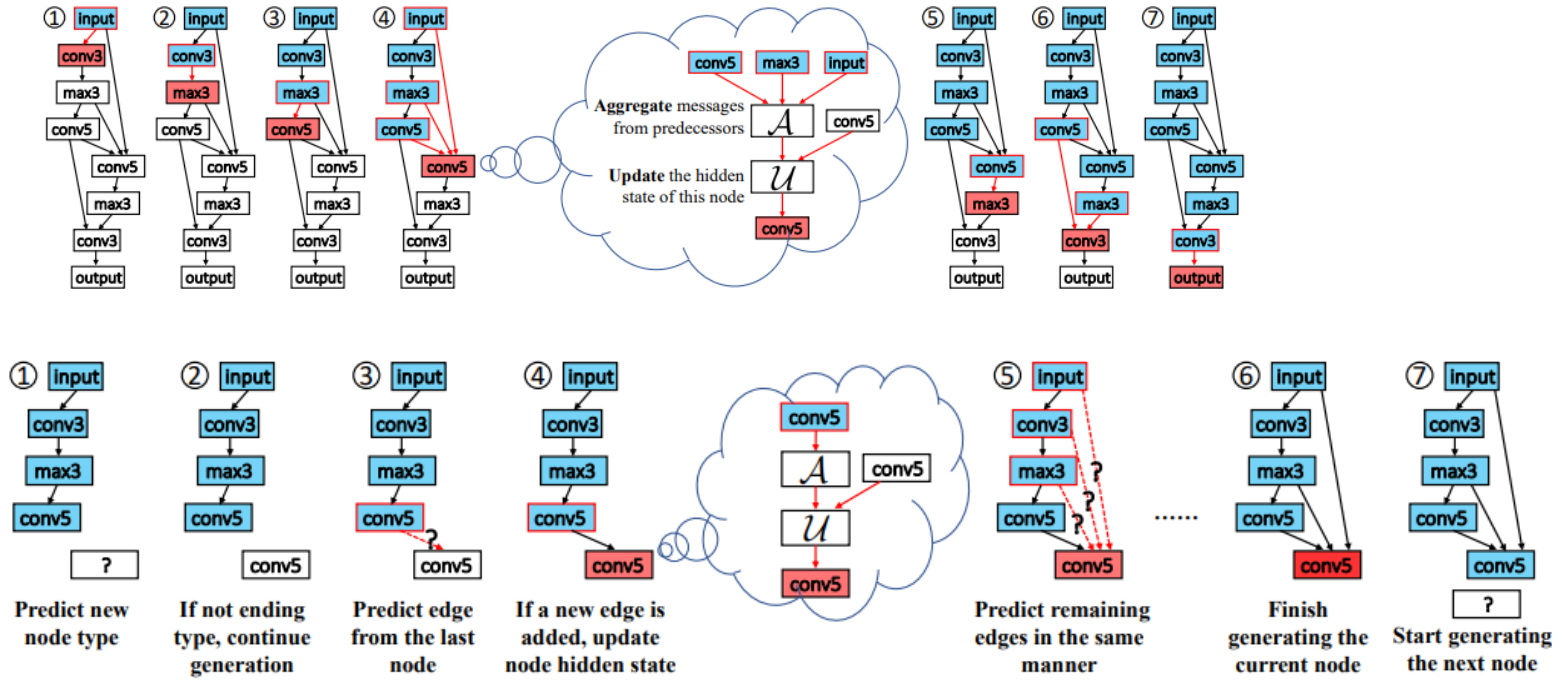


Figure 3: An illustration of the steps for generating a new node.

NIPS2019

也是编码图信息的

Semi-Implicit Graph Variational Auto-Encoders

Arman Hasanzadeh[†*], Ehsan Hajiramezanali[†*], Nick Duffield[†], Krishna Narayanan[†], Mingyuan Zhou[‡], Xiaoning Qian[†]

† Department of Electrical and Computer Engineering, Texas A&M University
{armanihm, ehsanr, duffieldng, krn, xqian}@tamu.edu
‡ McCombs School of Business, The University of Texas at Austin
mingyuan.zhou@mccombs.utexas.edu

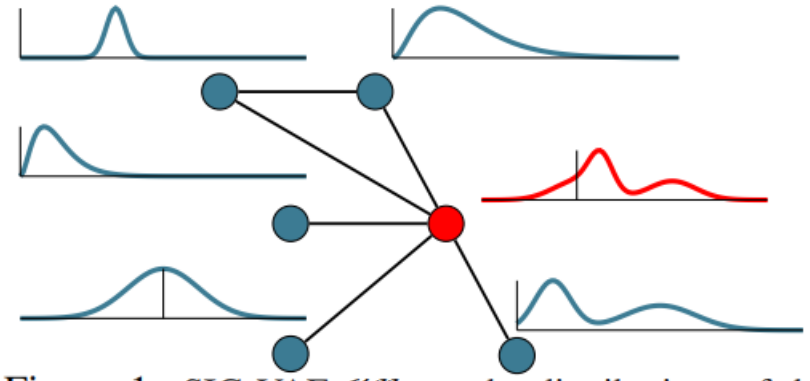


Figure 1: SIG-VAE diffuses the distributions of the neighboring nodes, which is more informative than sharing deterministic features, to infer each node's latent distribution.

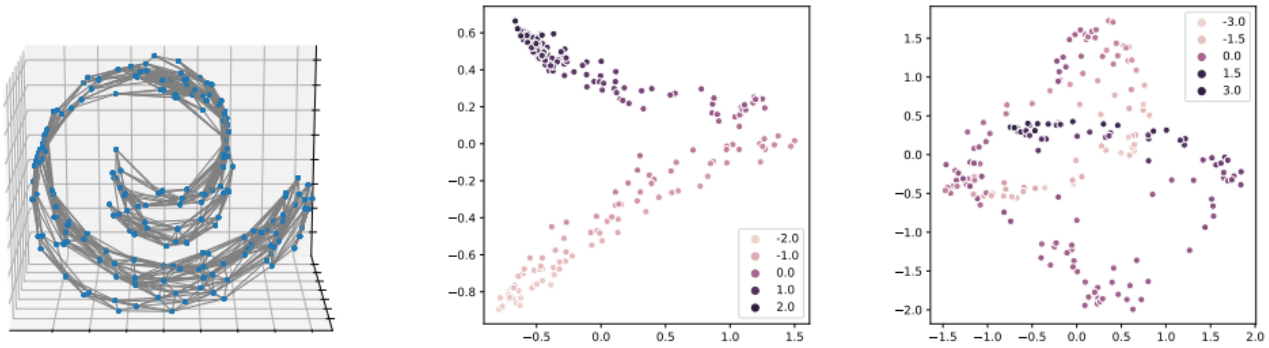


Figure 2: Swiss roll graph (left) and its latent representation using SIG-VAE (middle) and VGAE (right). The latent representations (middle and right) are heat maps in $\mathbb{R}^3$. We expect that the embedding of the Swiss roll graph with inner-product decoder to be a curved plane in $\mathbb{R}^3$, which is clearly captured better by SIG-VAE.

NIPS2020

编码图的成员之间的关系

Dirichlet Graph Variational Autoencoder

Jia Li[1], **Jianwei Yu**[1], **Jiajin Li**[1], **Honglei Zhang**[3], **Kangfei Zhao**[1],
Yu Rong[2], **Hong Cheng**[1], **Junzhou Huang**[2]
[1] The Chinese University of Hong Kong
[2] Tencent AI Lab
[3] Georgia Institute of Technology
{lijia,jwyu,jjli,kfzhao,hcheng}@se.cuhk.edu.hk, zhanghonglei@gatech.edu
yu.rong@hotmail.com, jzhuang@uta.edu

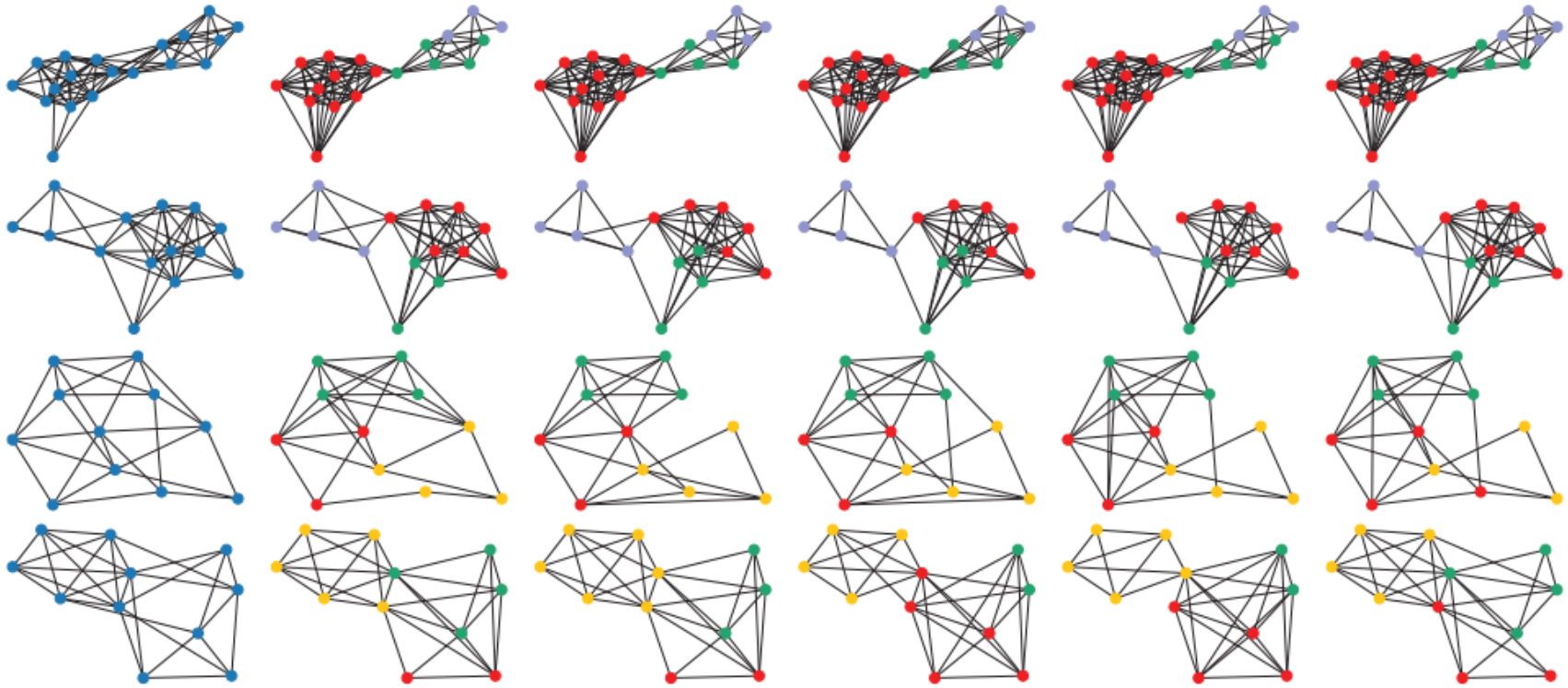


Figure 2: Left one in blue: the input graphs. Right five in colors: graph samples generated by DGVAE, where colors indicate latent cluster memberships with $K = 3$.

VAE性能优化

NIPS2019

工作：用庞加莱圆盘代替高斯分布，学习数据的分级结构
大扯黎曼几何，我人都傻了

Continuous Hierarchical Representations with Poincaré Variational Auto-Encoders

Emile Mathieu[†]
emile.mathieu@stats.ox.ac.uk

Charline Le Lan[†]
charline.lelan@stats.ox.ac.uk

Chris J. Maddison[†*]
cmaddis@stats.ox.ac.uk

Ryota Tomioka[‡]
ryoto@microsoft.com

Yee Whye Teh[†*]
y.w.teh@stats.ox.ac.uk

† Department of Statistics, University of Oxford, United Kingdom
∗ DeepMind, London, United Kingdom
‡ Microsoft Research, Cambridge, United Kingdom

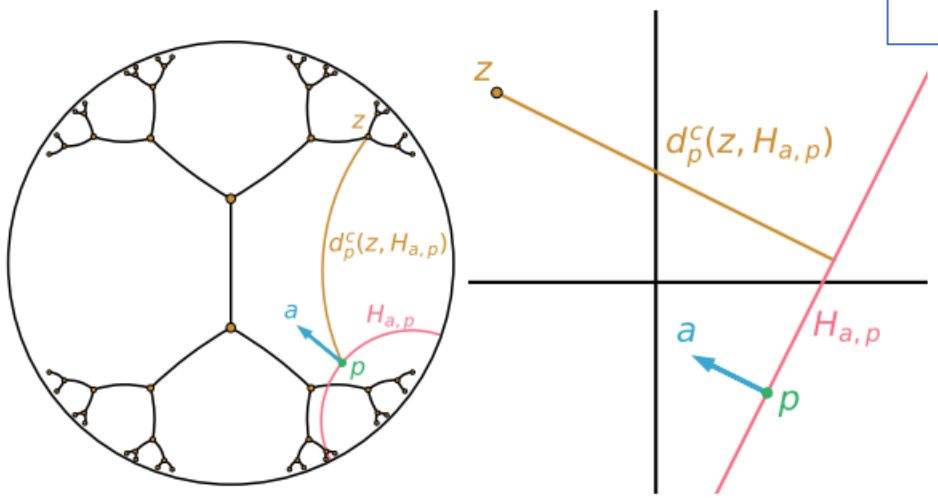
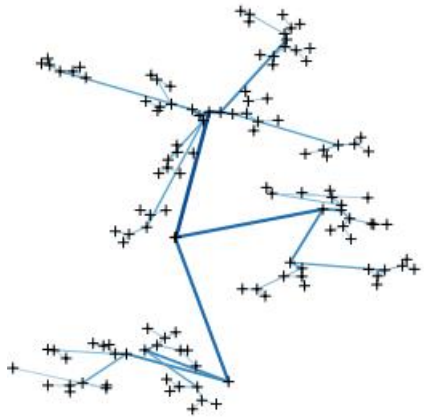


Figure 4: Illustration of an orthogonal projection on a hyperplane in a Poincaré disc (Left) and an Euclidean plane (Right).



NIPS2019

对伯努利VAE的损失函数进行修正，得到了更好的生成质量

这里的伯努利VAE、高斯VAE指的是：p(x|z)被我们认为是什么分布

The continuous Bernoulli: fixing a pervasive error in variational autoencoders

Gabriel Loaiza-Ganem
Department of Statistics
Columbia University
gl2480@columbia.edu

John P. Cunningham
Department of Statistics
Columbia University
jpc2181@columbia.edu

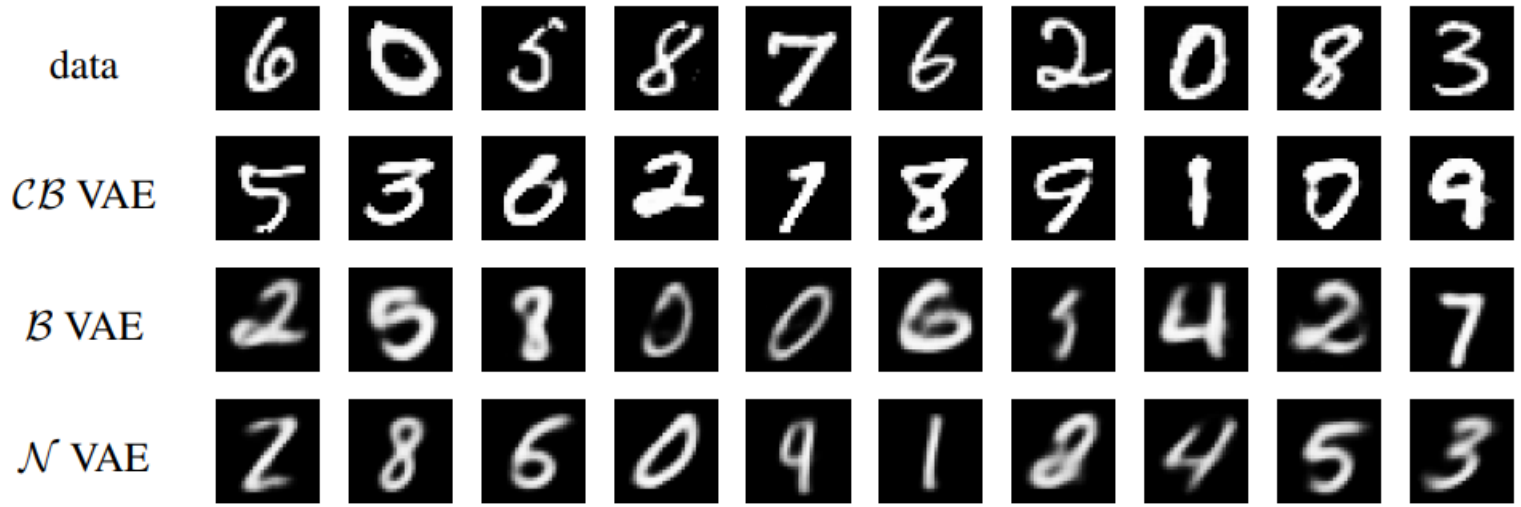


Figure 2: Samples from MNIST, continuous Bernoulli VAE, Bernoulli VAE, and Gaussian VAE.

Figure 2: Samples from MNIST, continuous Bernoulli VAE, Bernoulli VAE, and Gaussian VAE.

NIPS2020

动机：VAE的后验可能不准、可能方差太大
文章工作：在一个什么数据集上实验了各种目标函数对模型平均绝对误差的影响

Decision-Making with Auto-Encoding Variational Bayes

Romain Lopez[1], **Pierre Boyeau**[1], **Nir Yosef**[1,2,3], **Michael I. Jordan**[1,4], **and Jeffrey Regier**[5]

[1] Department of Electrical Engineering and Computer Sciences, University of California, Berkeley
[2] Chan-Zuckerberg Biohub, San Francisco
[3] Ragon Institute of MGH, MIT and Harvard
[4] Department of Statistics, University of California, Berkeley
[5] Department of Statistics, University of Michigan

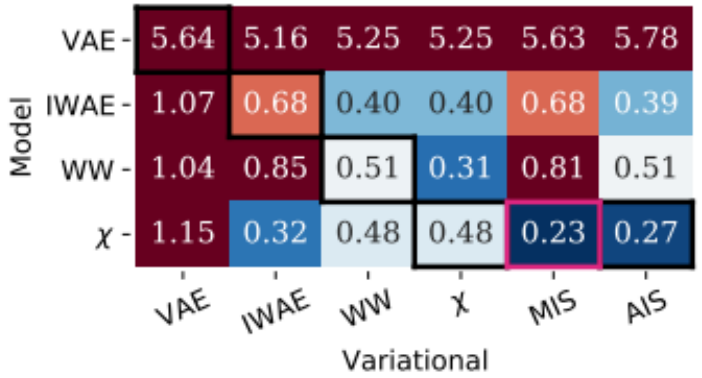
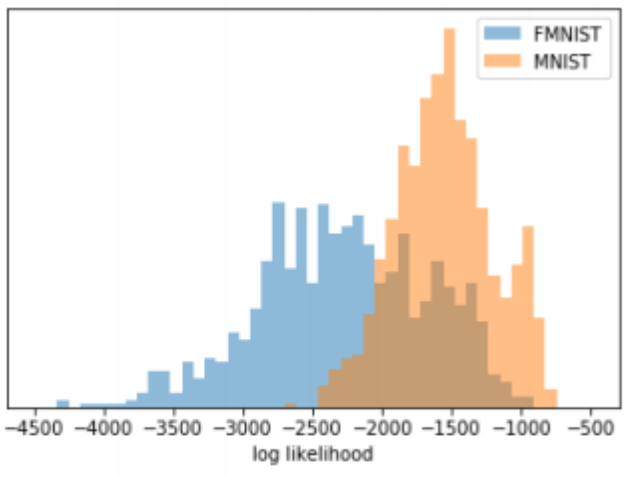


Figure 4: FDR MAE for scVI. Each row corresponds to an objective function for fitting the model parameters and each column corresponds to an objective function for fitting the variational parameters.

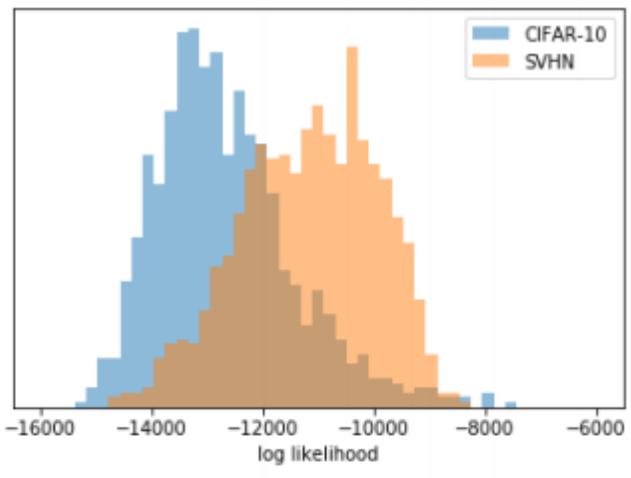
NIPS2020

Likelihood Regret: An Out-of-Distribution Detection Score For Variational Auto-encoder

动机：概率生成模型对某些不符合分
(out-of-distribution, ood)的样本给予
高概率，以前有方法解决这种问题，
是对VAE不管用，因此作者们就研究
一个出来

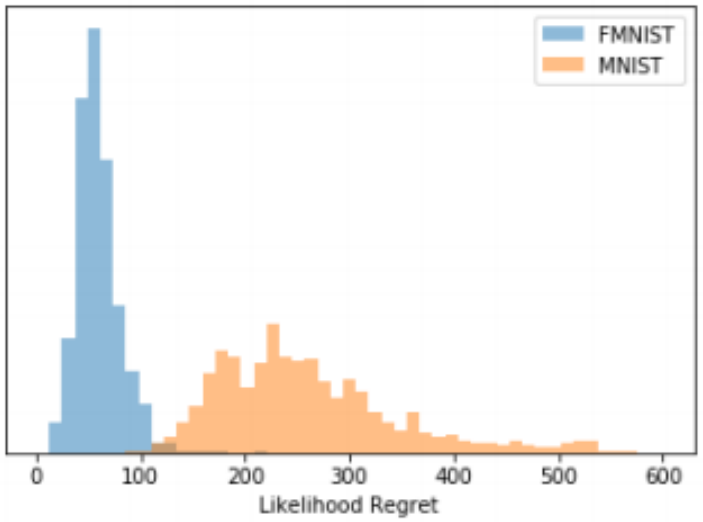


(a)

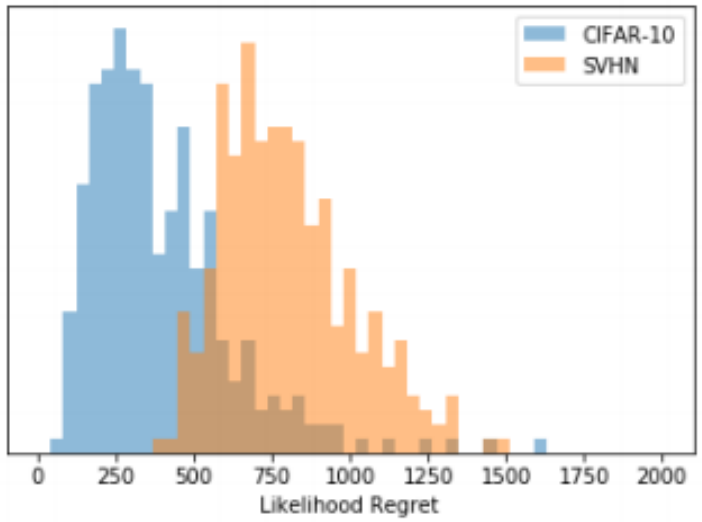


(b)

如右图：在左侧部分训练的
VAE能将右侧的图重构的很好，
这是我们不想要的。编码衣服
的编码器，就应该把数字编的
乱七八糟才行



(a)



(b)

he
AE
ed

Recursive Inference for Variational Autoencoders

NIPS2020

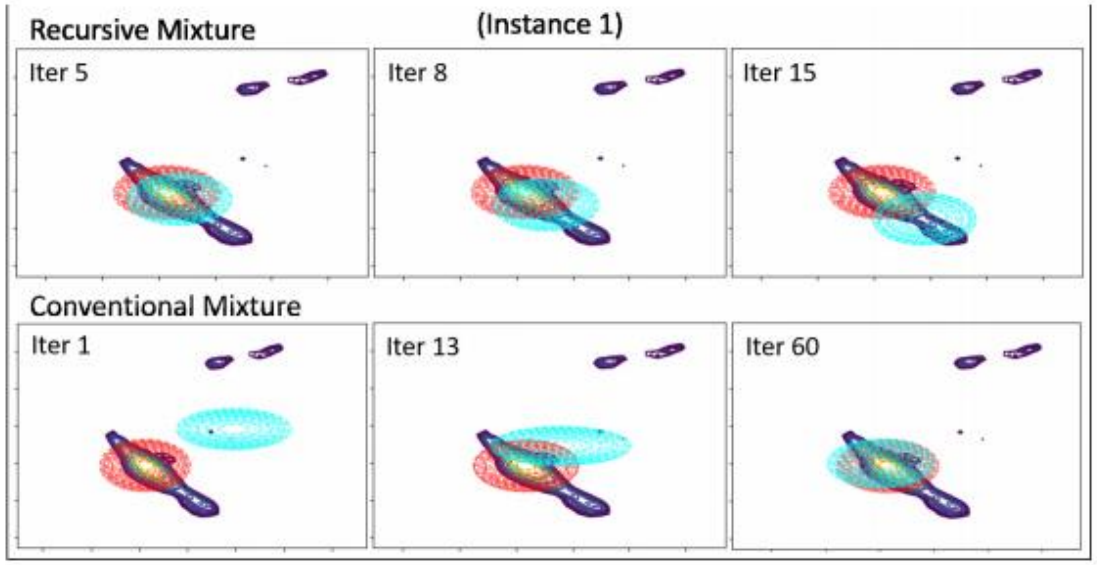
Minyoung Kim[1]
[1]Samsung AI Center
Cambridge, UK
mikim21@gmail.com

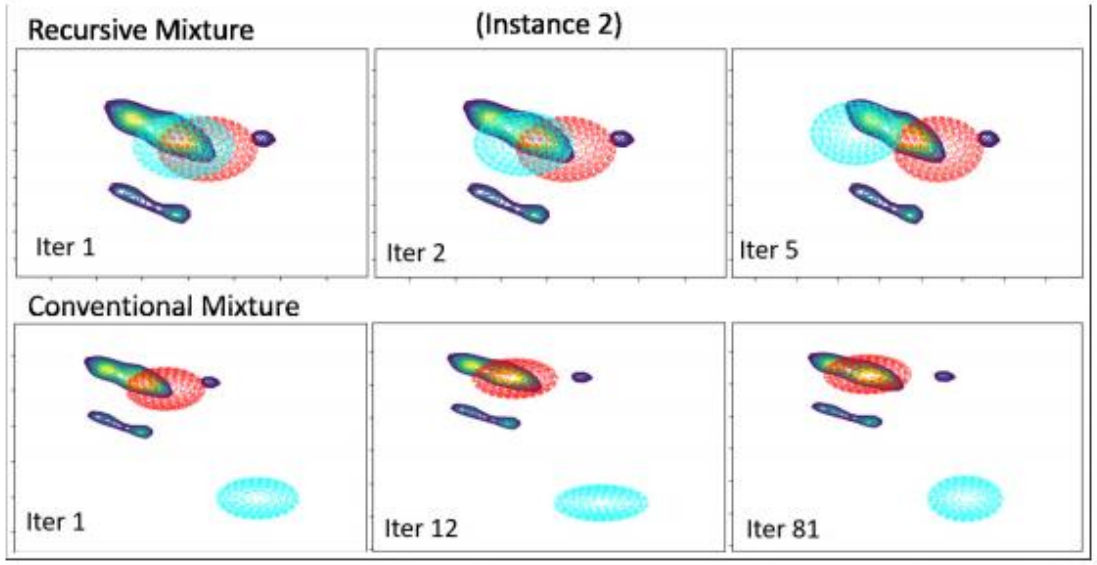
Vladimir Pavlovic[1,2]
[2]Rutgers University
Piscataway, NJ, USA
vladimir@cs.rutgers.edu

工作：为VAE提出新的优化算法

We propose a novel
recursive mixture estimation algorithm for VAEs that
iteratively augments the
current mixture with new components so as to maximally
reduce the divergence
between the variational and the true posteriors.

（我们提出了一种新的VAE递归混合估计算法，该算法用新的分量迭代地增加当前的混合，以最大限度地减小方差后验和真实后验之间的偏差。）





InfoVAE	two men are sitting on a couch while they wait for their parents to work. the woman is the world war. a man telling his hand. this man is sitting on bike. man jet ski is catching colorful waves. two little girl are happily not about to play twister.
CyclicalVAE	the men are ready to fight in the streets in front of the crowd with a sheet. a woman is standing with the canister in a band. a female age matured on the stage is looking in a microscope. a man is eating a small meal at a nightclub. the performer is taking his photos by a statue. there is a crowd of people, and several women posing for a picture.
LaggingVAE	this church choir sings to the masses as they sing joyous songs from the book at a church. this church choir sings to the masses as they sing joyous songs from the book at a church. this church choir sings to the masses as they sing joyous songs from the book at a church. a woman gets picture taken in front of the masses. a woman gets picture taken in front of the masses. a woman gets picture taken in front of the masses.
VAE-MINE	a choir including three people sing and dance on the stage in front of the masses. two musicians are playing the drums and a girl sits on a piano. a young family sits while waiting a girl at the bottom of a church. a man takes a photo of the girl. a little girl has a toy and a digital camera. the woman with red shirt is smiling to the girl.

Table 4: Sentences generated by interpolating between the encodings of "**this church choir sings to the masses as they sing joyous songs from the book at a church**." and "**a woman with a green headscarf, blue shirt and a very big grin**.".

EMNLP2019

Neural Gaussian Copula for Variational Autoencoder

Prince Zizhuang Wang
Department of Computer Science
University of California Santa Barbara
zizhuang_wang@ucsb.edu

William Yang Wang
Department of Computer Science
University of California Santa Barbara
william@cs.ucsb.edu

动机：经典VAE假设隐空间各个维度是独立的，而真实分布则有可能不是这样的

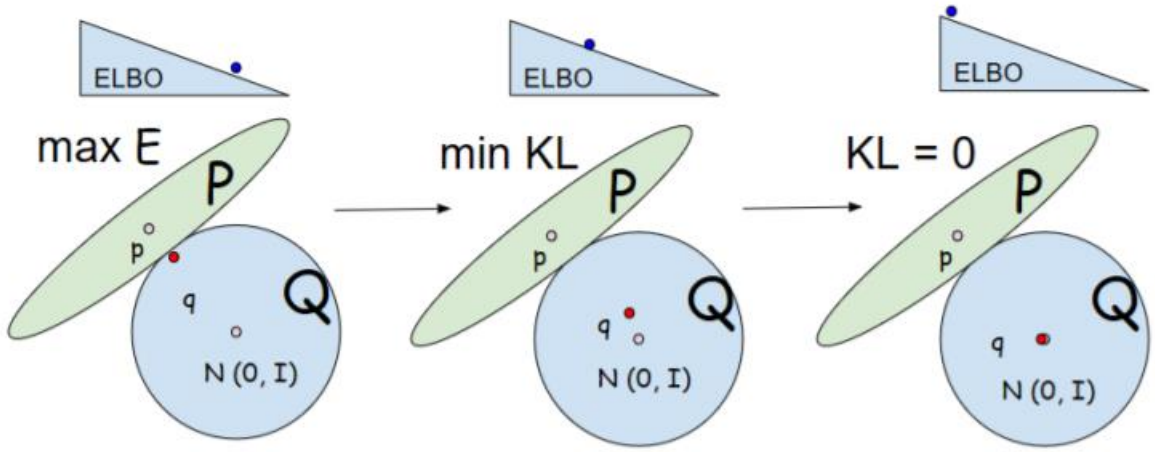


Figure 2: Training stage of VAE. Initially, the model tries to maximize ELBO by maximizing $\mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[p(\mathbf{x}|\mathbf{z})]$. Once $\mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[p(\mathbf{x}|\mathbf{z})]$ is maximized, the model maximizes ELBO by minimizing KL. During this stage, the posterior starts to move closer to the prior. In the final stage, the posterior collapses to the prior. But, the ELBO and $\mathbb{E}_{q(\mathbf{z}|\mathbf{x})}[p(\mathbf{x}|\mathbf{z})]$ are already maximized, which means the model keeps constraining KL and there are not enough gradients to move the posterior away from the prior anymore.

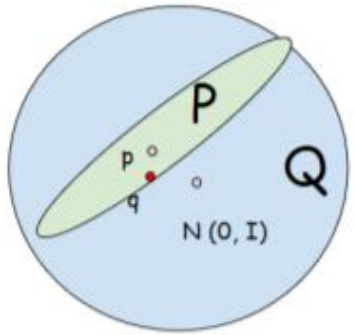
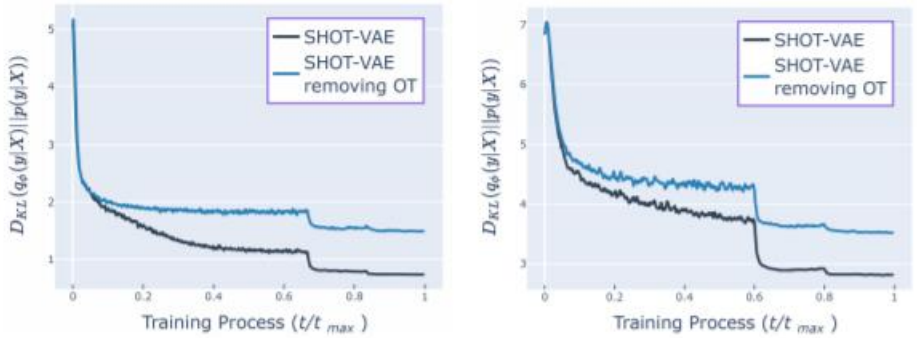


Figure 3: **Ideal** final stage of Copula-VAE. The family of distributions that contains the true posterior is now a subset of the variational family.

AAAI2021

任务：半监督学习图像分类
数据集：CIFAR-10、CIFAR-100
动机：1.有些数据集包含分类标签，ELBO不能利用数据集的标签信息
2.对ELBO的优化存在瓶颈

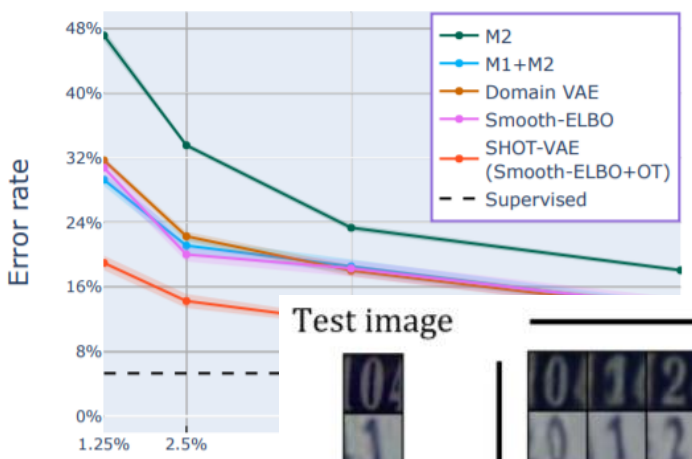
贡献：1.将分类loss加入ELBO中，提出了“Smooth-ELBO”
2.提出“OT-approximation”，解决ELBO优化瓶颈问题



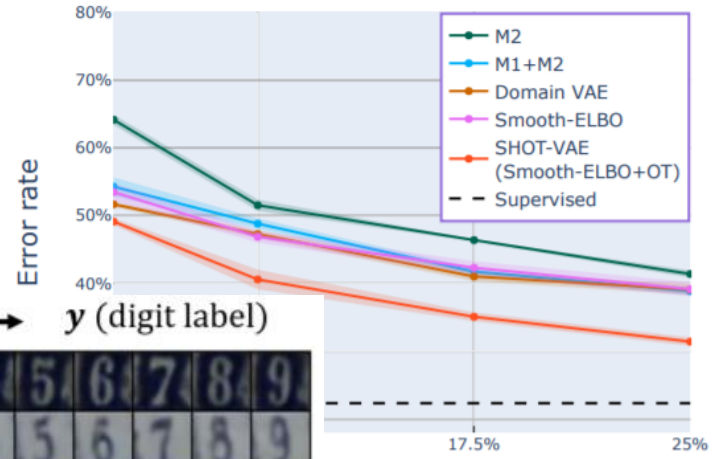
使用OT对半监督学习中的KLD项的影响

SHOT-VAE: Semi-supervised Deep Generative Models With Label-aware ELBO Approximations

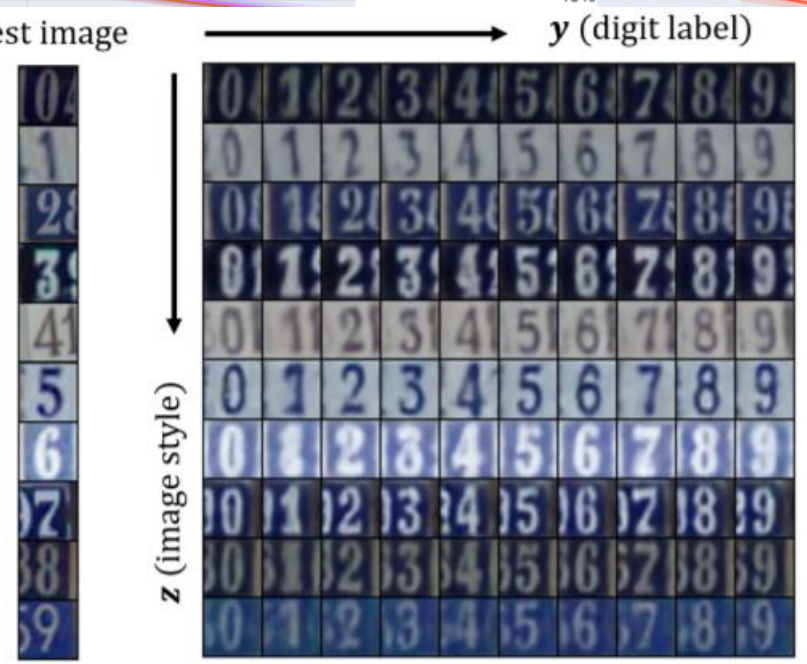
Hao-Zhe Feng,[1] Kezhi Kong,[2] Minghao Chen[1]
Tianye Zhang,[1] Minfeng Zhu,[1] Wei Chen[1]
[1] Zhejiang University
[2] University of Maryland, College Park



使用不



abeled ratio
分类的错误率



VAE部分生成效果

Few-shot learning

NIPS2020

The Autoencoding Variational Autoencoder

A. Taylan Cemgil
DeepMind

Sumedh Ghaisas
DeepMind

Krishnamurthy Dvijotham
DeepMind

Sven Gowal
DeepMind

Pushmeet Kohli
DeepMind

动机：对于有限的数据，即使是VAE目标的全局最小化也不足以编码出我们希望表示所具有的自然属性
由VAE的解码器生成的样本不被编码器映射到相应的表示

做法：套娃

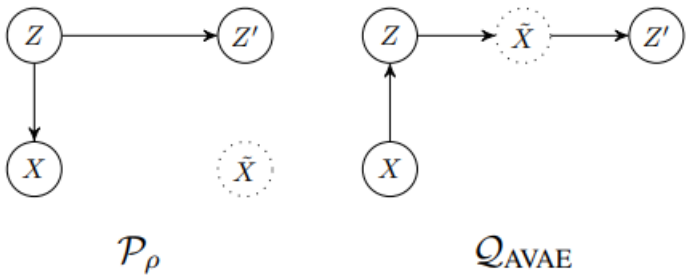


Figure 2: Graphical model of the extended target distribution $\mathcal{P}_\rho$, and the variational approximation $\mathcal{Q}_{\text{AVAE}}$. Here $\tilde{X}$ is a sample generated by the decoder that is subsequently encoded by the encoder.

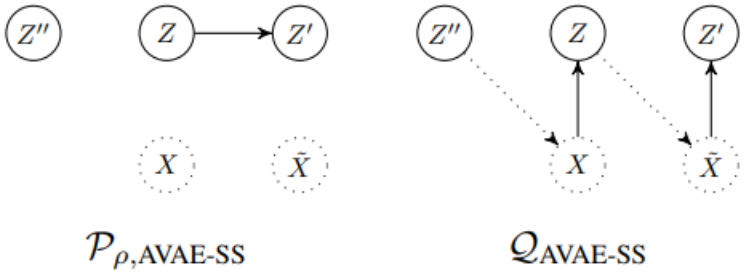


Figure 4: Graphical model of the AVAE-SS target distribution $\mathcal{P}_{\rho,\text{AVAE-SS}}$, and the variational approximation $\mathcal{Q}_{\text{AVAE-SS}}$. Here both X and $\tilde{X}$ is a sample generated by the decoder. The decoder factors (dotted arcs) is fixed (as pretrained by normal VAE).

解缠VAE

AAAI2021

Variational Disentanglement for Rare Event Modeling

Zidi Xiu, Chenyang Tao, Michael Gao, Connor Davis, Benjamin Goldstein, Ricardo Henao
Duke University
{zidi.xiu, chenyang.tao, michael.gao, connor.davis, benjamin.a.goldstein, ricardo.henao}@duke.edu

摘要：越来越多的丰富可用的医疗数据正与机器学习方法相结合，为临床决策系统创造了新的机会。然而在医疗风险预测应用中，我们感兴趣的病例（Rare Event，稀有事件）的比例相比于可用的病例通常非常低（也就是，训练集里不同标签的样本数非常不平衡）。基于这种动机，本文提出了一种变分解缠方法，半参数地学习严重不平衡问题中的西游事件。具体来说，本文利用潜在空间上的极端分布行为来从稀有事件中提取信息。，并且开发了一种结合了【generalized additive model】和【isotonic neural nets】的鲁棒的【prediction arm】

模型在各种real world dataset，包括新冠队列的死亡率预测方面都发挥了很好的效果

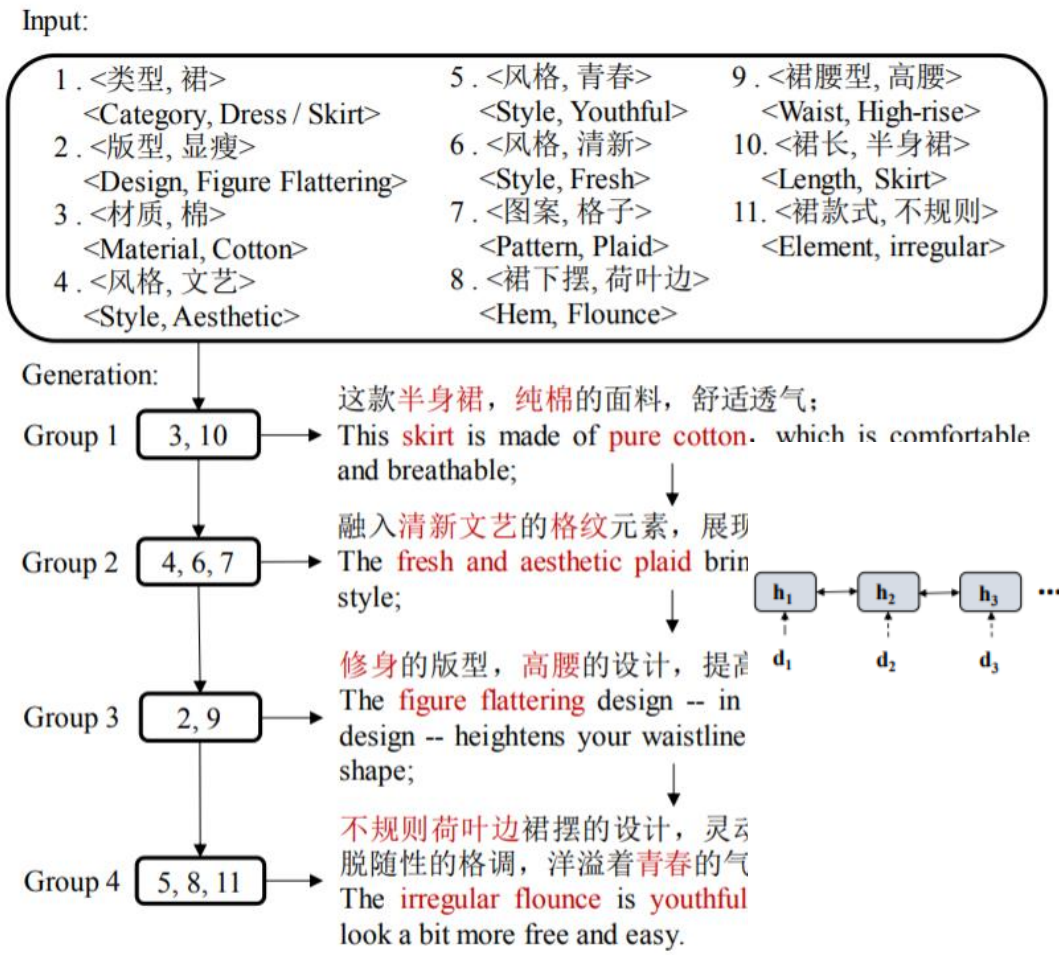
其他

EMNLP2019

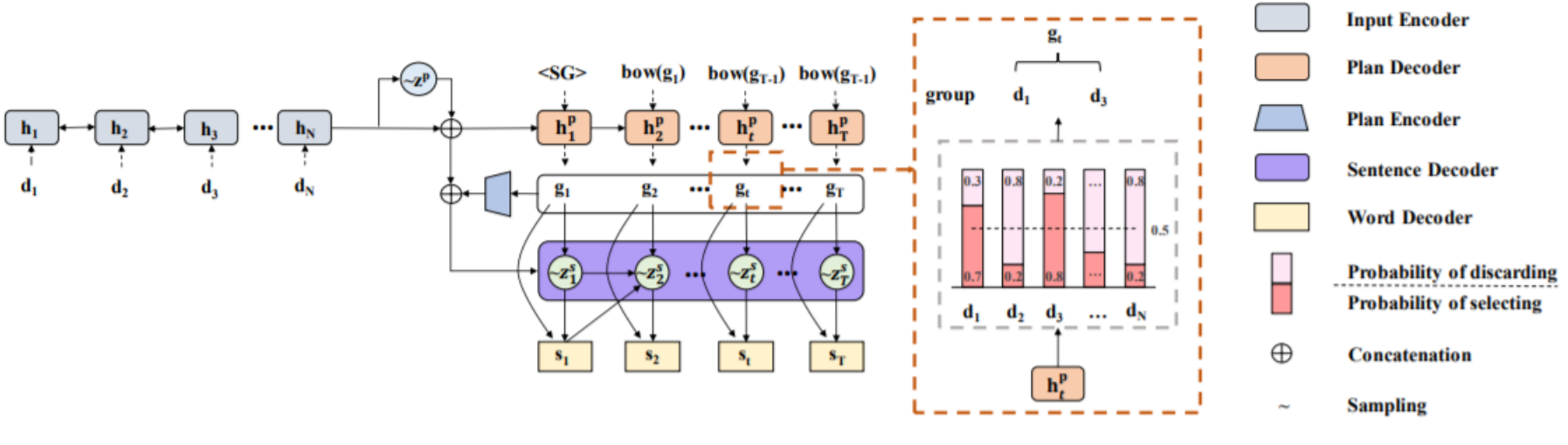
Long and Diverse Text Generation with Planning-based Hierarchical Variational Model

Zhihong Shao[1], Minlie Huang[1]*, Jiangtao Wen[1], Wenfei Xu[2], Xiaoyan Zhu[1]
[1] Institute for Artificial Intelligence, State Key Lab of Intelligent Technology and Systems
[1] Beijing National Research Center for Information Science and Technology
[1] Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
[2] Baozun, Shanghai, China
szh19@mails.tsinghua.edu.cn, aihuang@tsinghua.edu.cn
jtwen@tsinghua.edu.cn, xuwenfeilittle@gmail.com
zxy-dcs@tsinghua.edu.cn

任务：长文本生成（具体看图）
方法：分级生成



T：生成句子句子数
N：Input标签个数
先生成全局z，然后逐步生成各个句子的h，采样分组，生成各个句子的z，然后生成句子



EMNLP2020

Learning VAE-LDA Models with Rounded Reparameterization Trick

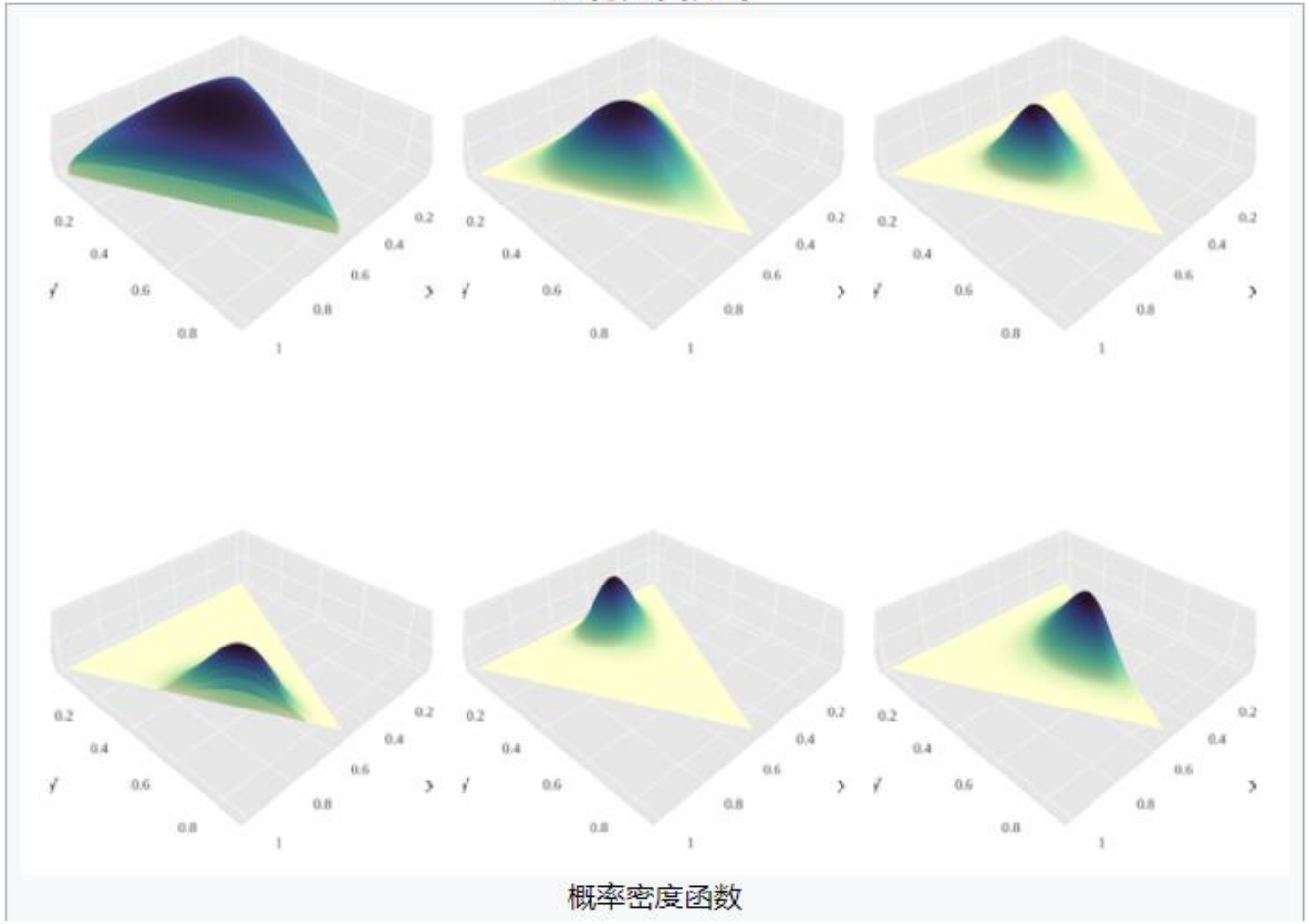
Runzhi Tian
School of EECS
University of Ottawa, Canada
rtian081@uottawa.ca

Yongyi Mao
School of EECS
University of Ottawa, Canada
ymao@uottawa.ca

Richong Zhang
BDBC and SKLSDE
Beihang University, China
zhangrc@act.buaa.edu.cn

动机：当样本被编码为迪利克雷分布时，原来的重参数化技巧就用不了了
贡献：本文作者想出了“四舍五入重参数化技巧”来解决这个问题

狄利克雷分布



概率密度函数

概率密度函数 [编辑]

维度$K \geq 2$的狄利克雷分布在参数$\alpha_1, \ldots, \alpha_K > 0$上、基于欧几里得空间数，定义为：

$$f(x_1, \ldots, x_K; \alpha_1, \ldots, \alpha_K) = \frac{1}{\mathrm{B}(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1}$$

EMNLP2020

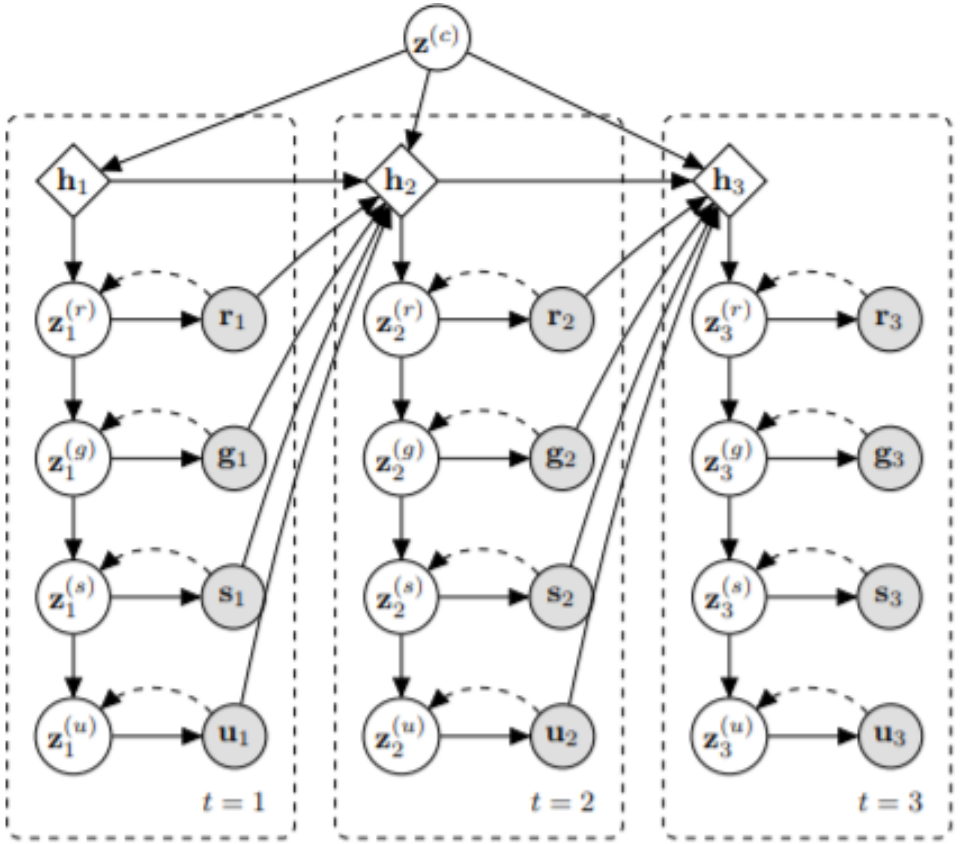


Figure 1: Graphical representation of VHDA. Solid and dashed arrows represent generation and recognition respectively.

Variational Hierarchical Dialog Autoencoder for Dialog State Tracking Data Augmentation

Kang Min Yoo[1] **Hanbit Lee**[1] **Franck Dernoncourt**[2] **Trung Bui**[2]
Walter Chang[2] **Sang-goo Lee**[1]

[1]Seoul National University, Seoul, Korea
[2]Adobe Research, San Jose, CA, USA
{kangminyoo,skcheon,sglee}@europa.snu.ac.kr
{dernonco,bui,wachang}@adobe.com

功能：生成用于DST的训练数据

r：语者信息
g：语者目标
s：dialogue state
u：语者说的一句话

如图所示，一次全部生成到位，和工读生说byebye!

AAAI2021

Learning Energy-Based Model with Variational Auto-Encoder as Amortized Sampler

Jianwen Xie, Zilong Zheng, Ping Li

Cognitive Computing Lab
Baidu Research
10900 NE 8th St. Bellevue, WA 98004, USA
{jianwenxie, v_zhengzilong, liping11}@baidu.com

用最大似然估计训练基于能量的模型（EBMs） 需要马尔科夫链蒙特卡罗（MCMC）抽样来逼近数据和模型分布之间KL散度的梯度，但是MCMC慢，而且MCMC和EBMs融合比较困难。因此本文采用VAE来抽样，供EBMs的训练使用。

先让MCMC教会VAE抽样的分布，再用VAE抽样来训练EBMs

参考资料

- b站VAE相关视频
- 熵、交叉熵、KL散度相关视频
- 提出VAE的论文
- 一篇讲VAE的blog
- 推导VAE代码中KL部分算式的论文