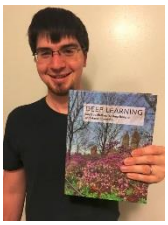
VAE

—— chenzhijun
BeiHang University
2019.6.20

深度生成模型就是利用神经网络来建模条件分布 p(x|z;θ)。

- 生成对抗网络（Generative Adversarial Network，GAN）[Ian Goodfellow, 2014]
- 变分自编码器（Variational Autoencoder，VAE）[Diederik P. Kingma, 2013]



Ian Goodfellow

Diederik P. Kingma

PhD 期间发表的论文包括 :VAE, Adam，还有 Variational Dropout

[PDF] Generative Adversarial Nets - NIPS Proceedings
https://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf ▾ 翻译此页
作者：I Goodfellow - 2014 - 被引用次数：8405 - 相关文章
Ian J. Goodfellow*, Jean Pouget-Abadie†, Mehdi Mirza, Bing Xu, David ... *Ian Goodfellow is now a research scientist at Google, but did this work earlier as a ...

Auto-Encoding Variational Bayes
https://arxiv.org › stat ▾ 翻译此页
作者：DP Kingma - 2013 - 被引用次数：4506 - 相关文章
2014年5月1日 - From: Diederik P Kingma M.Sc. [view email] [v1] Fri, 20 Dec 2013 20:58:10 UTC (3,884 KB) [v2] Mon, 23 Dec 2013 13:19:52 UTC (7,549 KB)

Adam: A Method for Stochastic Optimization
https://arxiv.org › cs ▾ 翻译此页
作者：DP Kingma - 2014 - 被引用次数：20744 - 相关文章
2014年12月22日 - From: Diederik P Kingma M.Sc. [view email] [v1] Mon, 22 Dec 2014 13:54:29 UTC (280 KB) [v2] Sat, 17 Jan 2015 20:26:06 UTC (283 KB)

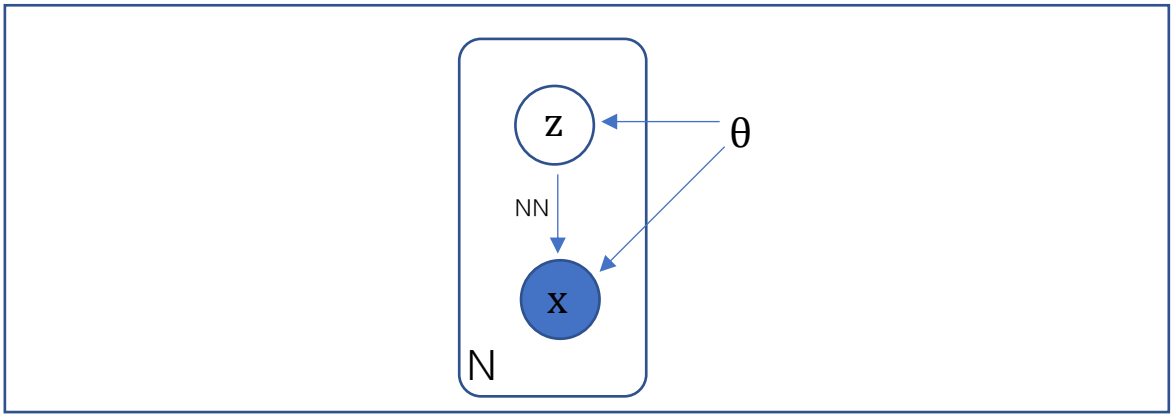
目录

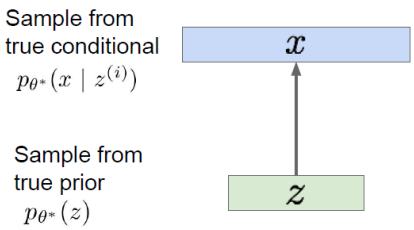
0. 面对的问题

前提：现在手里只有 N 个 datapoints(图片)： $\mathrm{X} = \{x^1, \ldots, x^N\}$ 。
初始求解目的：根据 data 对概率模型进行参数估计。求解出 datapoint x 的分布，能够生成无限多的新数据。(后面做完 VAE 之后，不仅完成了初始的目的，还能得到更多收获)。

1. 构造概率模型，提出数学假设

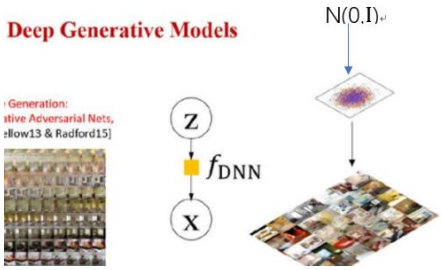
直接刻画 x 的概率分布 f(x)是困难的。
为了模型合理，模型便于求解，作者 Kingma 构建出概率模型：





其中，z 是低维 representation vector ，Z 的分布已被假设，Z~N(0,I)
$p_\theta(\mathrm{z})$=N(0,I)，$p_\theta(\mathrm{x}|\mathrm{z}) = \mathrm{N}(\mu_\theta(\mathrm{z}), \Sigma(\mathrm{z}))$

即，构造出带隐变量 z 的经典 **Deep Generative Model。**
此模型的思想，相似于**高斯混合模型 Gaussian Mixture Model (GMM)**，z 是类别（男女），x 是 observed data（身高）。



2. 求解概率模型

2.1 初试——EM 算法的失效

Data likelihood: $p_\theta(x) = \int p_\theta(z)p_\theta(x|z)dz$

Intractible to compute p(x|z) for every z!

Posterior density also intractable: $p_\theta(z|x) = p_\theta(x|z)p_\theta(z)/p_\theta(x)$

Intractable data likelihood

ML，概率统计中的经典问题——积分 intractable，posterior intractable 的问题。(或者，论文里经常说的不能表示成 closed-form 的形式，或者，can not be calculated analytically)

想象混合高斯情况下的 EM，没有面对以上的两个 intractable。因为 data likelihood 的边缘积分可以积分出来（实际上是累加），后验也是可求的，所以可以求解。但是面对现在的场景，EM 失效。

2.2 构造一个新东西——$q_\phi(z|x)$

构造 “recognition/inference network”，$q_\phi(z|x)$，使得$q_\phi(z|x)$能近似的逼近$p_\theta(x|z)$ 。

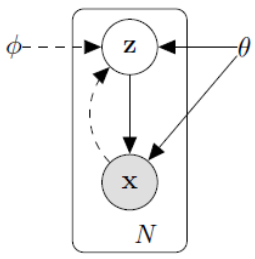
构造出来这个东西，才使得最终算法得解。

现在，我们把$p_\theta(x|z)$ 叫做“generation network”；把$q_\phi(z|x)$ 叫做“recognition/inference network”。

概率模型中，多了 parameters——ϕ。

当我们想求 datapoint x^i 对应的z^i 的概率分布情况，就可以用“recognition/inference network”来求解，即$q_\phi(z^i|x^i)$，也可写作$q(z^i; x^i, \phi)$。

所以，图示：



2.3 有了新东西后，推导 likelihood 的下届——为了得到 optimization objective

再次，现在所要面对的问题：参数估计，概率模型求解。

已有 observed data： $x^{(i)}$，i=1,2,⋯,N

构造的 probabilistic model 为： $\boldsymbol{p_\theta(\mathbf{z})}$**=N(0,I)，** $\boldsymbol{p_\theta(\mathbf{x}|\mathbf{z})} = \mathbf{N}(\boldsymbol{\mu_\theta(\mathbf{z}), \sigma^2_\theta(\mathbf{z})})$ **——Neural Network，**

待求的 model parameters： 其中 $\mu_\theta(z), \sigma^2_\theta(z)$中的参数θ待求；另外$\boldsymbol{\phi}$待求。

求解核心的思路：MLE 最大似然估计。

$log p_\theta(x^{(i)}) = \int_{z^{(i)}} q_\phi(z^{(i)}|x^{(i)}) log p_\theta(x^{(i)}) dz$ （$p_\theta(x^{(i)})$ does not depend on z）

$= \int_{z^{(i)}} q_\phi(z^{(i)}|x^{(i)}) log \frac{p_\theta(x^{(i)},z^{(i)})}{p_\theta(z^{(i)}|x^{(i)})} dz$ (Bayes Rule)

$= \int_{z^{(i)}} q_\phi(z^{(i)}|x^{(i)}) \log[\frac{p_\theta(x^{(i)},z^{(i)})}{q_\phi(z^{(i)}|x^{(i)})} \cdot \frac{q_\phi(z^{(i)}|x^{(i)})}{p_\theta(z^{(i)}|x^{(i)})}] dz$ (Multiply by constant)

$= \int_{z^{(i)}} q_\phi(z^{(i)}|x^{(i)}) \log \frac{p_\theta(x^{(i)},z^{(i)})}{q_\phi(z^{(i)}|x^{(i)})} dz + \int_{z^{(i)}} q_\phi(z^{(i)}|x^{(i)}) \log \frac{q_\phi(z^{(i)}|x^{(i)})}{p_\theta(z^{(i)}|x^{(i)})} dz$

$= \int_{z^{(i)}} q_\phi(z^{(i)}|x^{(i)}) \log \frac{p_\theta(x^{(i)},z^{(i)})}{q_\phi(z^{(i)}|x^{(i)})} dz + \mathbb{KL}(q_\phi(z^{(i)}|x^{(i)})||p_\theta(z^{(i)}|x^{(i)}))$

$ELBO$证据下届 Evidence Lower Bound: $L_b = \int_{z^{(i)}} q_\phi(z^{(i)}|x^{(i)}) \log \frac{p_\theta(x^{(i)},z^{(i)})}{q_\phi(z^{(i)}|x^{(i)})} dz$

即，总结以上推导，得到：

$log p_\theta(x^{(i)}) = L_b + \mathbb{KL}(q_\phi(z^{(i)}|x^{(i)})||p_\theta(z^{(i)}|x^{(i)}))$

在推导 EM 时，用 Jensen 不等式也是推导到了这一步。殊途同归。(用了不同的证明方法，得到同一结果)。

至此，在面对混合高斯模型 GMM 简单问题时，用 EM 已经 OK。如上所述，EM 失效。

继续推导：

$L_b = \int_{z^{(i)}} q_\phi(z^{(i)}|x^{(i)}) \log \frac{p_\theta(x^{(i)},z^{(i)})}{q_\phi(z^{(i)}|x^{(i)})} dz$

$= \int_{z^{(i)}} q_\phi(z^{(i)}|x^{(i)}) \log \frac{p_\theta(x^{(i)}|z^{(i)}) p_\theta(z^{(i)})}{q_\phi(z^{(i)}|x^{(i)})} dz$

$= \int_{z^{(i)}} q_\phi(z^{(i)}|x^{(i)}) \log p_\theta(x^{(i)}|z^{(i)}) dz + \int_{z^{(i)}} q_\phi(z^{(i)}|x^{(i)}) \log \frac{p_\theta(z^{(i)})}{q_\phi(z^{(i)}|x^{(i)})} dz$

$= \int_{z^{(i)}} q_\phi(z^{(i)}|x^{(i)}) \log p_\theta(x^{(i)}|z^{(i)}) dz - \mathbb{KL}(q_\phi(z^{(i)}|x^{(i)})||p_\theta(z^{(i)}))$

$= \mathbb{E}_{z\sim q_\phi(z^{(i)}|x^{(i)})}[\log p_\theta(x^{(i)}|z^{(i)})] - \mathbb{KL}(q_\phi(z^{(i)}|x^{(i)})||p_\theta(z^{(i)}))$

上式的含义可以这么理解：

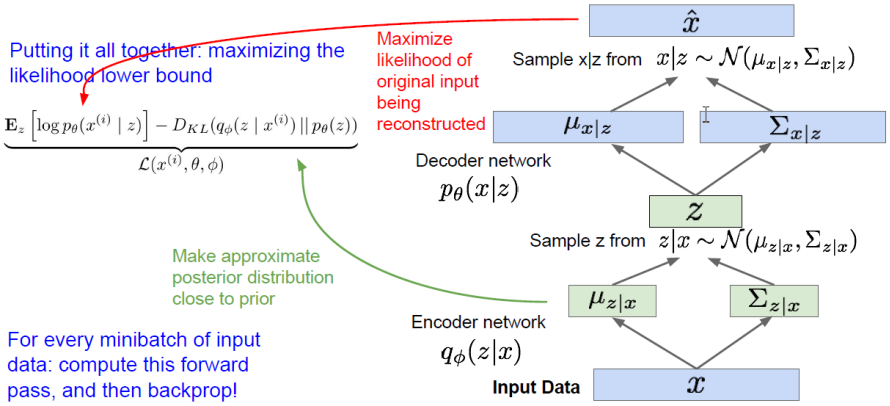
为了使得L_b尽可能的大，要$\mathbb{E}_{z\sim q_\phi(z^{(i)}|x^{(i)})}[\log p_\theta(x^{(i)}|z^{(i)})]$部分尽可能大，即每个样本重构出来的要像原来的 datapoint x； 另外要让$\mathbb{KL}(q_\phi(z^{(i)}|x^{(i)})||p_\theta(z^{(i)}))$部分尽可能小，即每个样本的 z 变量的分布靠近于 N(0,I)，z 变量不能胡来，它得受先验 N(0,I)的影响。

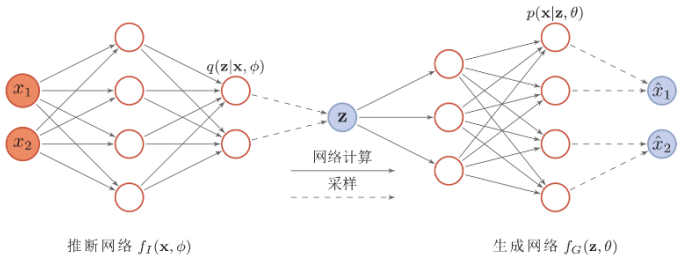
类比于 ML 中的许多算法（MAP 最大后验估计），可以理解为：

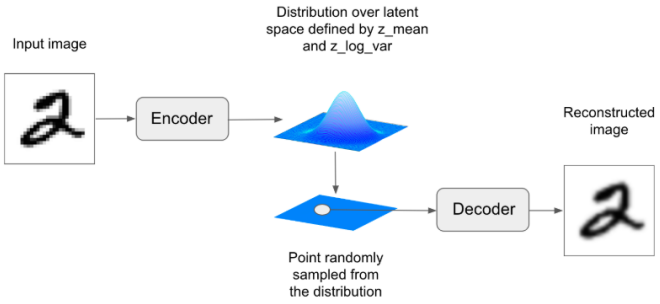
总体 Loss = Reconstruction Loss+ Regularization Loss

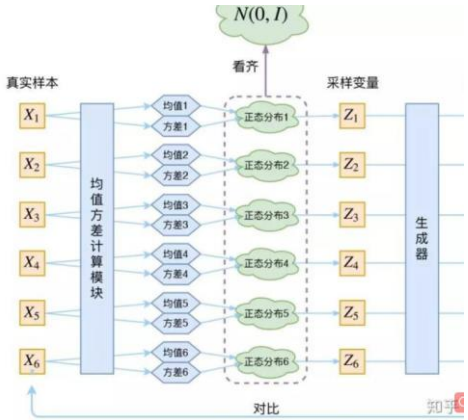
在 ELBO 的表达式中，有$q_\phi(z^{(i)}|x^{(i)})$部分，有$p_\theta(x^{(i)}|z^{(i)})$部分，如果我们现在非要画画，画一个图，那么这个图可以是如下这样的，为了表达清楚，用了几张看起来不同的图，但是图的意思都一样。

至此，和自编码器的结构不期而至：









我个人认为，至此，VAE 和自编码器在形式上很相近，并且都有信息压缩，信息放大的意思。

这么类比，也可以帮助人快速理解 VAE 的大体意思，而减小对 VAE 数学上的关注。

但是我认为，在推导 VAE，解释 VAE 的时候，即便不引入自编码器的结构，不说这个事情，也是没有任何

问题的，因为 VAE 没有用到自编码器所独有的数学理论上的东西。
所以：这些图在脑子里，能让自己对 VAE 中随机变量的结构有一个感性的认识，对 VAE 中信息压缩与信息放大有一个感性的认识。但具体到问题，具体到数学细节，还是看 VAE 本身，自编码器帮不上忙。

2.4 Optimization

以上，得到：

$$log p_\theta(x^{(i)}) = L_b + \mathbb{KL}(q_\phi(z^{(i)}|x^{(i)})||p_\theta(z^{(i)}|x^{(i)}))$$

ELBO: $L_b = \int_{z^{(i)}} q_\phi(z^{(i)}|x^{(i)}) \log \frac{p_\theta(x^{(i)},z^{(i)})}{q_\phi(z^{(i)}|x^{(i)})} dz$

$$= \int_{z^{(i)}} q_\phi(z^{(i)}|x^{(i)}) \log p_\theta(x^{(i)}|z^{(i)})\, dz \; - \; \mathbb{KL}(q_\phi(z^{(i)}|x^{(i)})||p_\theta(z^{(i)}))$$

$$= \mathbb{E}_{z\sim q_\phi(z^{(i)}|x^{(i)})}[\log p_\theta(x^{(i)}|z^{(i)})] \; - \; \mathbb{KL}(q_\phi(z^{(i)}|x^{(i)})||p_\theta(z^{(i)}))$$

从本质上说，在有了这个式子后，有数据$x^{(1)},\dots x^{(N)}$，初始化模型参数$\boldsymbol{\theta}$，$\boldsymbol{\phi}$，可以通过基于 gradient 的方法来做 optimization。
但是其中有两个困难，再加上两个 trick 之后，我们就可以用 gradient 的方法来做 optimization。
两个 trick：蒙特卡洛近似，重参数化技巧（reparameterization trick）。

两个困难，与相应的两个 trick:

“However, calculating the expectation and its gradients is non-trivial, often intractable”.

1.期望项难求，积分求不出来——用蒙特卡洛近似：

首先，ELBO中的KL项$\mathbb{KL}(q_\phi(z^{(i)}|x^{(i)})||p_\theta(z^{(i)})$有closed-form，可以计算出来。
（在 VAE 原文中，$q_\phi(z^{(i)}|x^{(i)})$被假设 z 的各个维度的值是独立的，在 z 的每个维度上，$-\mathrm{KL}(q(z|x;\phi)\|p(z;\theta)) = \frac{1}{2}(1+\log\sigma^2-\mu^2-\sigma^2)$）

其次，期望项$\mathbb{E}_{z\sim q_\phi(z^{(i)}|x^{(i)})}[\log p_\theta(x^{(i)}|z^{(i)})]$，积分 intractable，即无法积分成关于θ，$\boldsymbol{\phi}$的函数，L(θ, $\boldsymbol{\phi}$; $x^{(i)}$)

用采样的方法（蒙特卡洛方法）来做。

采样法（Sampling Method），也叫蒙特卡罗方法（Monte Carlo Method）,MC estimate，是 20 世纪 40 年代中期提出的一种通过随机采样的方法来近似估计一些计算问题的数值解

蒙特卡罗方法的基本思想可以归结为根据一个已知概率密度函数为$p(x)$的分布来计算函数$f(x)$的期望

$$\mathbb{E}[f(x)] = \int_x f(x)p(x)dx. \qquad (11.42)$$

当$p(x)$比较复杂时，很难用解析的方法来计算这个期望。为了计算$\mathbb{E}[f(x)]$，我们可以通过**数值解法**的方法来近似计算。首先从$p(x)$中独立抽取的N个样本$x^{(1)},x^{(2)},\cdots,x^{(N)}$，$f(x)$的期望可以用这$N$个样本的均值$\hat{f}_N$来近似。

$$\hat{f}_N = \frac{1}{N}\left(f(x^{(1)})+\cdots+f(x^{(N)})\right). \qquad (11.43)$$

根据大数定律，当N趋向于无穷大时，样本均值收敛于期望值。

$$\hat{f}_N \xrightarrow{P} \mathbb{E}_p[f(x)] \quad \text{当 } N\to\infty. \qquad (11.44)$$

这就是蒙特卡罗方法的理论依据。

$$\mathbb{E}_{z\sim q_\phi(z^{(i)}|x^{(i)})}[\log p_\theta(x^{(i)}|z^{(i)})] \approx \frac{1}{L}\sum_{l=1}^{L} \log p_\theta(x^{(i)}|z^{(i,l)})$$

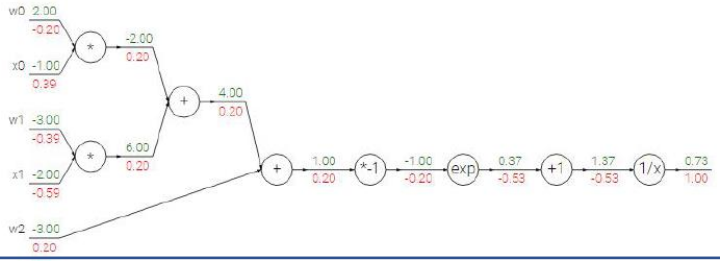
称$\frac{1}{L}\sum_{l=1}^{L}\log p_\theta(x^{(i)}|z^{(i,l)})$为$\mathbb{E}_{z\sim q_\phi(z^{(i)}|x^{(i)})}[\log p_\theta(x^{(i)}|z^{(i)})]$的estimator，估计量。作者给这个estimator叫作SGVB。

2.无法对随机变量进行求导，求 gradient——重参数化技巧（reparameterization trick）：

反向传播：

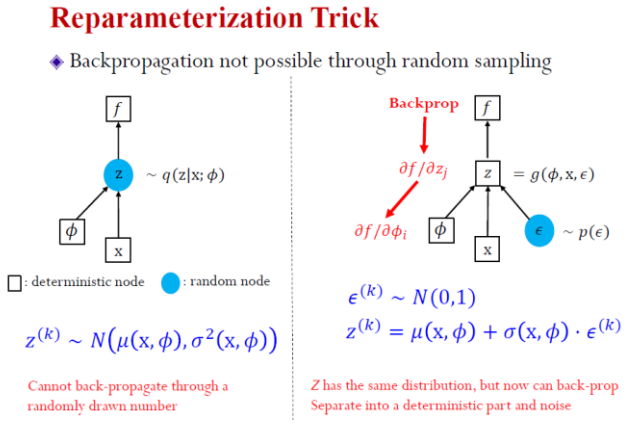
现在看看一个表达式：

$$f(w,x)=\frac{1}{1+e^{-(w_0x_0+w_1x_1+w_2)}}$$



VAE的model parameters为θ，$\boldsymbol{\phi}$，给模型输入datapoint x，它就能得到z的分布并采样出一个z，也得到最终的optimization objective,即ELBO。

得到z，是model计算中的一环。但是和上面的普通的反向传播不同的是，z是一个随机变量的一个节点，不是通过固定的变换而得到的一个固定的值。



所以，总体的VAE算法：

Algorithm 1 Minibatch version of the Auto-Encoding VB (AEVB) algorithm. Either of the two SGVB estimators in section 2.3 can be used. We use settings $M = 100$ and $L = 1$ in experiments.

$\theta, \phi \leftarrow$ Initialize parameters
repeat
  $\mathbf{X}^M \leftarrow$ Random minibatch of M datapoints (drawn from full dataset)
  $\epsilon \leftarrow$ Random samples from noise distribution $p(\epsilon)$
  $\mathbf{g} \leftarrow \nabla_{\theta,\phi}\tilde{\mathcal{L}}^M(\theta, \phi; \mathbf{X}^M, \epsilon)$ (Gradients of minibatch estimator (8))
  $\theta, \phi \leftarrow$ Update parameters using gradients g (e.g. SGD or Adagrad [DHS10])
until convergence of parameters (θ, ϕ)
return θ, ϕ

2.5 VAE 的用途

若得到了θ，$\boldsymbol{\phi}$，则可以：

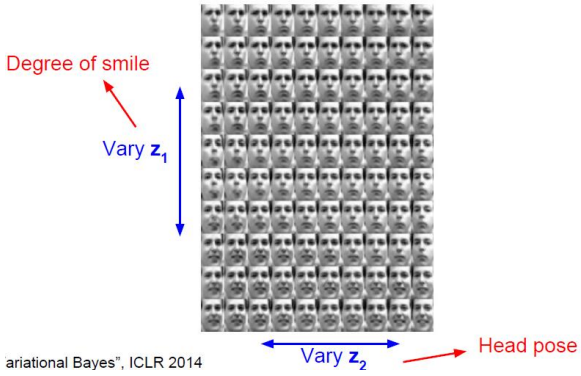
1. 生成。

   $p_\theta(x,z) = p_\theta(z) \cdot p_\theta(x|z)$　　Z~N(0,I)已有，$p_\theta(x|z)$已求得，可生成无限多新数据。

2. 分类。

   一方面，手头上 N 个 datapoint 的对应的 N 个隐变量分布得知，可知 N 个隐变量各自最可能的取值（手写体识别，知道每个图片写的是数字几）。另一方面，如果再给我一个新的 datapoint，可以利用“recognition/inference network”，$q_\phi(z|x)$”推理出此 datapoint 对应的隐变量分布，即可能取值信息。

其中，在第一个功能——生成时，有意思的是，可以调节，控制生成：



另外，在 NLP，CV 中也有很多应用，但是远远没有 GAN 火热。

3. References and code

[1] Kingma D P, Welling M. Auto-encoding variational bayes[J]. arXiv preprint arXiv:1312.6114, 2013.

[2] CS231 2017 https://www.bilibili.com/video/av19305792/?p=2

[3] 朱军 统计机器学习课件

[4] Tutorial on Variational Autoencoders ArXiv'2016

[5] 复旦大学邱锡鹏 《神经网络与深度学习》 https://nndl.github.io/

[6] Code(已跑通):
https://github.com/bojone/vae
http://dpkingma.com/?page_id=393