NICE: Non-linear Independent Components Estimation

Laurent Dinh, David Krueger, Yoshua Bengio

Flow Model: NICE

- 生成模型的本质，就是希望用一个我们知道的概率模型来拟合所给的数据样本，也就是说，我们得写出一个带参数θ的分布$q_\theta(x)$。
- 对于连续型的简单分布，我们也就只能写出高斯分布了，而且很多时候为了方便处理，我们只能写出各分量独立的高斯分布，这显然只是众多连续分布中极小的一部分，显然是不够用的。为了解决这个困境，我们通过积分来创造更多的分布：

$$q_\theta(x) = \int q(z) q_\theta(x|z) dz$$

  - 这里$q(z)$一般是标准的高斯分布，而$q_\theta(x|z)$可以选择任意的条件高斯分布（VAE）或者狄拉克分布（GAN）。这样的积分形式可以形成很多复杂的分布。理论上来讲，它能拟合任意分布。

Flow Model: NICE

- 现在分布形式有了，我们需要求出参数θ，那一般就是最大似然，假设真实数据分布为$\tilde{p}(x)$，那么我们就需要最大化目标:

$$\mathbb{E}_{x\sim\tilde{p}(x)}[\log q_{\theta}(x)]$$

- 然而$q_{\theta}(x)$是积分形式的，优化目标的计算会很困难。

- 为了解决这个问题，VAE设法去优化一个更强的上界，从而得到一个近似模型。GAN则是通过一个交替训练的方法绕开了这个困难。

Flow Model: NICE

- FLOW 模型则选择直接把这个积分算出来。

  - 具体来说，flow模型选择$q_\theta(x|z)$为狄拉克分布$\delta(x-g_\theta(z))$，而且$g_\theta(z)$必须是可逆的，也就是说:

$$x = g_\theta(z) \quad \leftrightarrow \quad z = f_\theta(x)$$

  - 要求z和x的维度一样。假设f, g的形式都知道了，那么优化目标的计算就可以通过对$q(z)$做一个积分变换$z = f(x)$得到。即本来有：

$$q(z) = \frac{1}{(2\pi)^{D/2}}\exp(-\frac{1}{2}\|z\|^2)$$

  - 现在做积分变换$z = f(x)$。但要注意：概率密度函数的变量代换并不是简单地将z替换为$f(x)$就行了，还多出了一个“雅可比行列式”的绝对值。也就是：

$$q(x) = \frac{1}{(2\pi)^{D/2}}\exp(-\frac{1}{2}\|f(x)\|^2)\left|\det\left[\frac{\partial f}{\partial x}\right]\right|$$

Flow Model: NICE

- 优化目标变为：

$$\log q(x) = -\frac{D}{2}\log(2\pi) - \frac{1}{2}\|f(x)\|^2 + \log\left|\det\left[\frac{\partial f}{\partial x}\right]\right|$$

- 通过观察这个新的优化目标，我们对f有两个要求：
  - ✓ 可逆，并且逆函数易于求解；（ f 的逆函数g就是我们希望得到的生成模型）
  - ✓ 对应的雅可比行列式要容易计算；

- 满足以上要求的情况下，这个优化目标是可以求解的。并且由于f容易求逆，因此一旦训练完成，我们就可以随机采样一个z，然后通过f的逆来生成一个样本$x = f^{-1}(z) = g(z)$，这就得到了生成模型。

Flow Model: NICE

- 相对而言，行列式的计算要比函数求逆要困难。我们知道，三角阵的行列式最容易计算，所以我们应该要想办法使得变换 f 的雅可比矩阵为三角阵。NICE的做法很精巧，它将D维的x分为两部分x_1, x_2，然后取下述变换：

$$h_1 = x_1$$
$$h_2 = x_2 + m(x_1)$$

- 其中x_1, x_2是x的某种划分，m是x_1的任意函数。也就是说，将x分为两部分，然后按照上述公式进行变换，得到新的变量h，这个被称为“加性耦合层”(Additive Coupling)。不失一般性，可以将x各个维度进行重排，使得$x_1 = x_{1:d}$为前d个元素，$x_2 = x_{d+1:D}$为$d + 1 \sim D$个元素。

Flow Model: NICE

- 不难看出，这个变换的雅可比矩阵$\left[\frac{\partial h}{\partial x}\right]$是一个三角阵，而且对角线全部为1，用分块矩阵表示为:

$$\left[\frac{\partial h}{\partial x}\right] = \begin{pmatrix} \mathbb{I}_d & \mathbb{O} \\ \left[\frac{\partial m}{\partial x_1}\right] & \mathbb{I}_{d:D} \end{pmatrix}$$

- 变换h的雅可比行列式为1，其对数值为0。这样就解决了行列式的计算问题。
- 同时，这个变换也是可逆的：

$$x_1 = h_1$$

$$x_2 = h_2 - m(h_1)$$

Flow Model: NICE

- 要注意到，变换h的第一部分是恒等变换。因此单个变换不能达到非常强的非线性，所以需要多个简单变换的复合，以达到强非线性，增强拟合能力：

$$x = h^{(0)} \leftrightarrow h^{(1)} \leftrightarrow h^{(2)} \leftrightarrow \ \dots \ \leftrightarrow h^{(n-1)} \leftrightarrow h^{(n)} = z$$

- 其中每个变换都是加性耦合层。这就好比流一般，积少成多，所以这样的一个流程称为一个“流（flow）”。也就是说，一个flow是多个耦合层的耦合。

- 由链式法则：

$$\left[\frac{\partial z}{\partial x}\right] = \left[\frac{\partial h^{(n)}}{\partial h^{(0)}}\right] = \left[\frac{\partial h^{(n)}}{\partial h^{(n-1)}}\right]\left[\frac{\partial h^{(n-1)}}{\partial h^{(n-2)}}\right]\cdots\left[\frac{\partial h^{(1)}}{\partial h^{(0)}}\right]$$

$$det\left[\frac{\partial z}{\partial x}\right] = det\left[\frac{\partial h^{(n)}}{\partial h^{(n-1)}}\right] det\left[\frac{\partial h^{(n-1)}}{\partial h^{(n-2)}}\right] \dots det\left[\frac{\partial h^{(1)}}{\partial h^{(0)}}\right] = 1$$

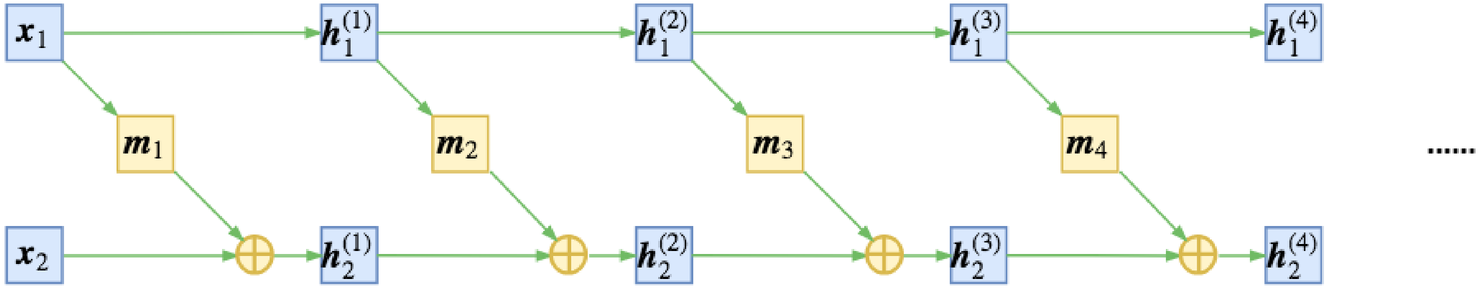
Flow Model: NICE

- 要注意，如果耦合的顺序一直保持不变，即:

$$h_1^{(1)} = x_1 \qquad h_1^{(2)} = h_1^{(1)}$$

$$h_2^{(1)} = x_2 + m_1(x_1) \qquad h_2^{(2)} = h_2^{(1)} + m_2\left(h_1^{(1)}\right) \ \ldots$$

- 那么最后还是$z_1 = x_1$，第一部分依然是初始输入，如下图：



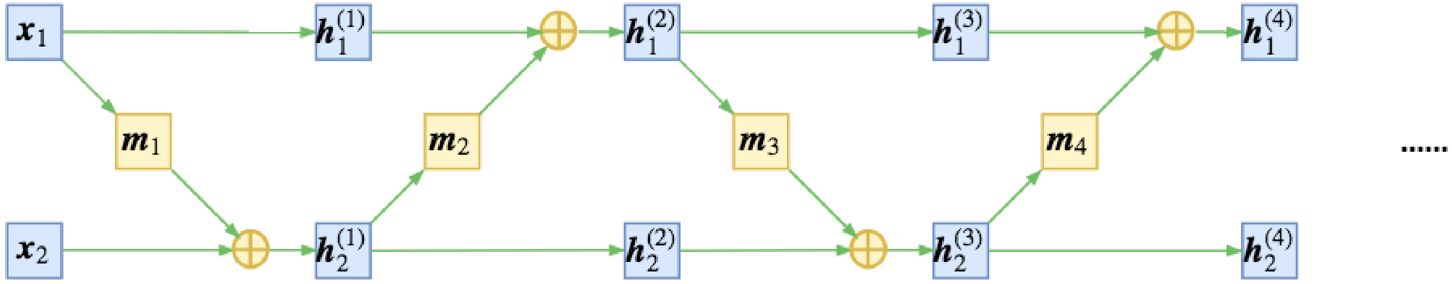
Flow Model: NICE

- 为了得到不恒等的变换，考虑在每次进行加性耦合前，打乱或反转输入的各个维度的顺序，或者简单地直接交换这两部分的位置，使得信息可以充分混合，比如：

$$h_1^{(1)} = x_1 \qquad h_1^{(2)} = h_1^{(1)} + m_2\left(h_2^{(1)}\right)$$

$$h_2^{(1)} = x_2 + m_1(x_1) \qquad h_2^{(2)} = h_2^{(1)} \quad \dots$$

- 如下图：



Flow Model: NICE

- Flow是基于可逆变换的模型，当模型训练完成之后，我们同时得到了一个生成模型和一个编码模型。随机变量z和输入样本x具有同一大小。当我们指定z为高斯分布时，它是遍布整个D维空间的，D也就是输入x的尺寸。但虽然x具有D维，但它未必就真正能遍布整个D维空间，比如MNIST图像虽然有784个像素，但有些像素不管在训练集还是测试集，都一直保持为0，这说明它远远没有784维那么大。

  - 也就是说，flow这种基于可逆变换的模型，天生就存在比较严重的维度浪费问题：输入数据明明都不是D维流形，但却要编码为一个D维流形。

Flow Model: NICE

- 为了解决这个情况，NICE引入了一个尺度变换层，它对最后编码出来的每个维度的特征都做了个尺度变换，也就是$z = s \otimes h^{(n)}$这样的形式，其中$s = (s_1, s_2, \ldots, s_D)$也是一个要优化的参数向量（各个元素非负）。
- 这个s向量能识别该维度的重要程度（越小越重要，越大说明这个维度越不重要，接近可以忽略），起到压缩流形的作用。注意这个尺度变换层的雅可比行列式就不再是1了，可以算得它的雅可比矩阵为对角阵:

$$\left[\frac{\partial z}{\partial h^{(n)}}\right] = diag(s)$$

- 它的行列式为：$\prod_i s_i$，有对数似然：

$$\log q(x) \sim -\frac{1}{2}\|s \otimes f(x)\|^2 + \sum_i \log s_i$$

Flow Model: NICE

- 其实这个尺度变换层可以换一种更加清晰的方式描述——带参数方差的正态分布：

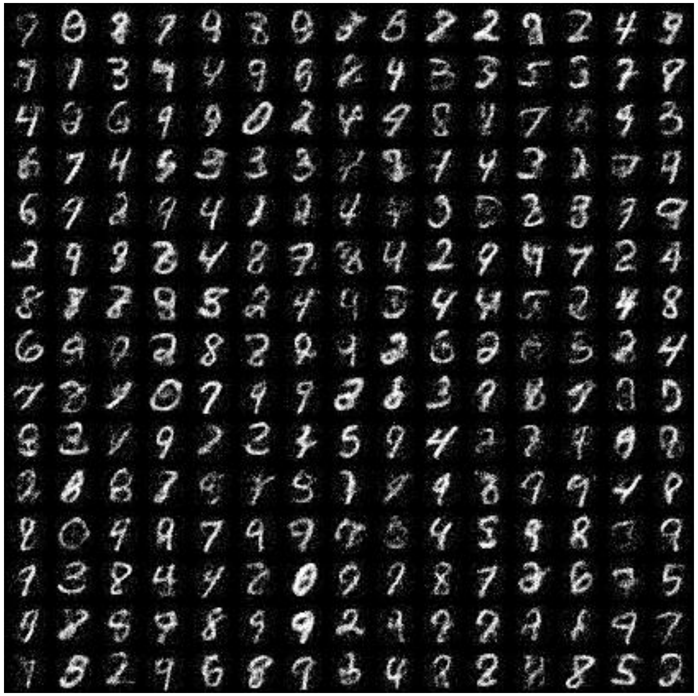
$$q(z) = \frac{1}{(2\pi)^{D/2}\prod_{i=1}^{D}\sigma_i}\exp(-\frac{1}{2}\sum_{i=1}^{D}\frac{z_i^2}{\sigma_i^2})$$

- 代入$z = f(x)$的变换，取对数得：

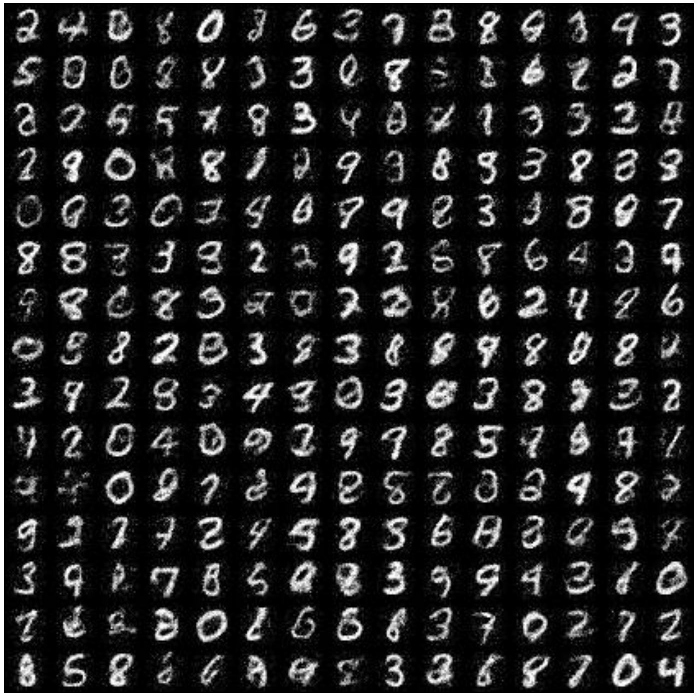
$$\log q(x) \sim -\frac{1}{2}\sum_{i=1}^{D}\frac{f_i^2(x)}{\sigma_i^2} - \sum_i \log \sigma_i$$

  - 两式做一个对比，就有$s_i = {}^1/_{\sigma_i}$。所以尺度变换层等价于将先验分布的方差也作为训练参数，如果方差足够小，就可以认为该维度所表示的流形坍缩为一个点，从而总体流形的维度减1，暗含了降维的可能。

Flow Model: NICE



无噪声训练



有噪声训练

Flow Model: NICE

- 模型细节：

- 加性耦合层需要将输入分为两部分，NICE采用交错分区，即下标为偶数的作为第一部分，下标为奇数的作为第二部分，而每个$m(x)$则简单地用多层全连接（5个隐藏层，每个层1000节点，ReLU激活）。在NICE中一共耦合了4个加性耦合层；
- NICE的模型还是比较庞大的，按照上述模型，模型的参数量约为 2×10^7 ，也就是两千万的参数只为训练一个MNIST生成模型。模型参数还是过于庞大；
- 且非线性变换的部分只能使用全连接层的结构。