10 May 能量视角下的GAN模型（三）：生成模型=能量模型

By 苏剑林 | 2019-05-10 | 20787位读者

本文的模型在ImageNet(128x128)上的条件生成效果

今天要介绍的结果还是跟能量模型相关，来自论文《Implicit Generation and Generalization in Energy-Based Models》。当然，它已经跟GAN没有什么关系了，但是跟本系列第二篇所介绍的能量模型关系较大，所以还是把它放到这个系列好了。

我当初留意到这篇论文，是因为机器之心的报导《MIT本科学神重启基于能量的生成模型，新框架堪比GAN》，但是说实在的，这篇文章没什么意思，说句不中听的，就是炒冷饭系列，媒体的标题也算中肯，是“重启”。这篇文章就是指出能量模型实际上就是某个特定的Langevin方程的静态解，然后就用这个Langevin方程来实现采样，有了采样过程也就可以完成能量模型的训练，这些理论都是现成的，所以这个过程我在学习随机微分方程的时候都想过，我相信很多人也都想过。因此，我觉得作者的贡献就是把这个直白的想法通过一系列炼丹技巧实现了。

但不管怎样，能训练出来也是一件很不错的事情，另外对于之前没了解过相关内容的读者来说，这确实也算是一个不错的能量模型案例，所以我论文的整体思路整理一下，让读者能够更全面地理解能量模型。

能量分布

跟《能量视角下的GAN模型（二）：GAN = “分析” + “采样”》一样，假设我们有一批数据$x_1, x_2, \dots, x_n \sim p(x)$，我们希望用一个概率模型去拟合它，我们选取的模型为

$$q_\theta(x) = \frac{e^{-U_\theta(x)}}{Z_\theta} \tag{1}$$

其中U_θ是带参数θ的未定函数，我们称为“能量函数”，而Z_θ是归一化因子（配分函数）

$$Z_\theta = \int e^{-U_\theta(x)} dx \tag{2}$$

这样的分布可以称为“能量分布”，在物理中也被称为“玻尔兹曼分布”。

为了求出参数θ，我们先定义对数似然函数：

$$\mathbb{E}_{x\sim p(x)}\big[\log q_\theta(x)\big] \tag{3}$$

我们希望它越大越好，也就是希望

$$L_\theta = \mathbb{E}_{x\sim p(x)}\big[-\log q_\theta(x)\big] \tag{4}$$

越小越好，为此，我们对L_θ使用梯度下降。我们有（具体推导参考第二篇）

$$\nabla_\theta \log q_\theta(x) = -\nabla_\theta U_\theta(x) + \mathbb{E}_{x\sim q_\theta(x)}\left[\nabla_\theta U_\theta(x)\right] \tag{5}$$

所以

$$\nabla_\theta L_\theta = \mathbb{E}_{x\sim p(x)}\left[\nabla_\theta U_\theta(x)\right] - \mathbb{E}_{x\sim q_\theta(x)}\left[\nabla_\theta U_\theta(x)\right] \tag{6}$$

这意味着梯度下降的更新公式是

$$\theta \leftarrow \theta - \varepsilon\Big(\mathbb{E}_{x\sim p(x)}\left[\nabla_\theta U_\theta(x)\right] - \mathbb{E}_{x\sim q_\theta(x)}\left[\nabla_\theta U_\theta(x)\right]\Big) \tag{7}$$

Langevin方程

在式(6)中，$\mathbb{E}_{x\sim p(x)}\left[\nabla_\theta U_\theta(x)\right]$是容易估算的，直接抽样一批真实数据来计算就行了；但是$\mathbb{E}_{x\sim q_\theta(x)}\left[\nabla_\theta U_\theta(x)\right]$却很困难，因为我们不知道怎么实现从$q_\theta(x)$中采样。

《能量视角下的GAN模型（二）：GAN = “分析” + “采样”》中的思路是定义另外一个容易采样的分布$q_\varphi(x)$，然后改为从$q_\varphi(x)$中采样，同时去缩小$q_\varphi(x)$和$q_\theta(x)$的差异，使得$q_\varphi(x)$确实可以成为$q_\theta(x)$的一个良好近似。但这篇论文不一样，它直接从能量模型对应的Langevin方程采样。

其实思路很简单，在上一篇文章已经已经提到过，对于Langevin方程：

$$x_{t+1} = x_t - \frac{1}{2}\varepsilon\nabla_x U(x_t) + \sqrt{\varepsilon}\alpha,\quad \alpha\sim\mathcal{N}(\alpha;0,1) \tag{8}$$

当$\varepsilon\to 0$且$t\to\infty$时，序列$\{x_t\}$所服从的分布就是$q_\theta(x)$，换句话说，$q_\theta(x)$是该Langevin方程的静态分布，再换句话说，那就是给定$U_\theta(x)$后（$q_\theta(x)$也确定了），式(8)的递归过程就可以帮我们得到一批从$q_\theta(x)$采样的样本来。

嗯嗯，有了这个采样过程，那就完事了呀，首先$\mathbb{E}_{x\sim q_\theta(x)}\left[\nabla_\theta U_\theta(x)\right]$可以估算了，因此能量模型可以完成训练了；训练完成之后，还是由式(8)帮助我们从中采样出一批新样本了，这样就完成生成过程了。

模型细节

当然，理论是这样子，实际操作肯定有很多细节，而且少不了炼丹。我当初也就只是思考到这一步，觉得里边的边角问题太多，难以解决，就没有继续做下去了。但作者坚持下去了，终究是跑通了，这一点我是很佩服的。

首先是作者往模型$U_\theta(x)$加入了谱归一化，而$U_\theta(x)$本身就相当于GAN中的判别器地位，所以加入谱归一化是可以理解的。其次，在训练的过程中，用的能量函数不是$U_\theta(x)$，而是加上一个小的L2正则：$U_\theta(x) + \lambda U_\theta^2(x)$，其中$\lambda$是一个小的正常数，作者的意思是这样会使得整个loss更光滑，训练起来更稳定（使用时还是$U_\theta(x)$）。

然后，回到采样问题，采样是通过式(8)进行的，它是一个迭代过程，既然是迭代就需要初始值。然而如果直接从随机分布（比如均匀分布）中采样随机向量作为初始值，作者提到会出现模式单一的问题，即迭代出来的图片形式比较单一，导致采样不充分，所以作者维护了一个Buffer，它把历史的的采样结果缓存起来，作为下一次采样的候选初始值。

总的来说，模型的更新过程如下：

> 假定数据样本分布为$p(x)$，选定迭代步长ε（参考值为1/200）、迭代步数K（参考值20～50）和batch size N，Buffer记为$\mathcal{B}$，初始化是空集。
>
> 循环执行，直到收敛：
>
> > 循环执行，得到一批真假样本：
> >
> > 1、从$p(x)$中采样一个真样本x_r，加入到当前批；
> >
> > 2、以95%的概率从$\mathcal{B}$（或者以5%的概率从均匀分布）选取一个样本作为初始值$x_{f,0}$；
> >
> > 3、以$x_{f,0}$为初始值，迭代式(8)共K步，得到$x_{f,K}$；
> >
> > 4、将$x_{f,K}$作为假样本x_f，加入到当前批，同时加入到$\mathcal{B}$。
> >
> > 有了真假样本后，执行一步优化器，优化目标为：
> >
> > $$\frac{1}{N}\sum_{x_r,x_f}\Big\{U_\theta(x_r) - U_\theta(x_f) + \lambda\big[U_\theta^2(x_r) - U_\theta^2(x_f)\big]\Big\}$$

而训练完成后的采样，同样需要维护Buffer，并且作者为了保证多样性，他将模型分别训练几次，得到若干个不同权重的统一模型，然后同时从这若干个模型中采样，并且共享、共同维护一个Buffer。其他细节问题大家直接看原论文即可，因为不打算复现，所以就不考究了。

作者实现：https://github.com/openai/ebm_code_release

个人总结

总的来说，我认为这是一篇中规中矩、差强人意的论文。首先思路和理论都是成熟的，能量模型与Langevin方程的关系前人早已得出，不算什么创新；但是能够攻克其中的细节难题，真正把这个思路落实下去，也不是一件容易的事情，体现了作者在生成模型领域深厚的（炼丹）功底。从能量模型的角度看，也可以说是为训练复杂的能量模型提供了一个可行的方案。

至于效果上，可以说它媲美GAN，也可以说比不上GAN。作者主要在Cifar10和ImageNet上做实验，这两个数据集当然很难，可以说一般的GAN都生成不好，从效果图来看，确实可以PK大多数GAN了，在Cifar10上明显完胜Glow。说它比不上，则是感觉它太有技巧性了，不够优雅，比如Langevin方程的所导致的采样思路，我感觉就没有什么底，维护一个Buffer的做法，虽然实践效果还可以，但显然工程味道太浓了....

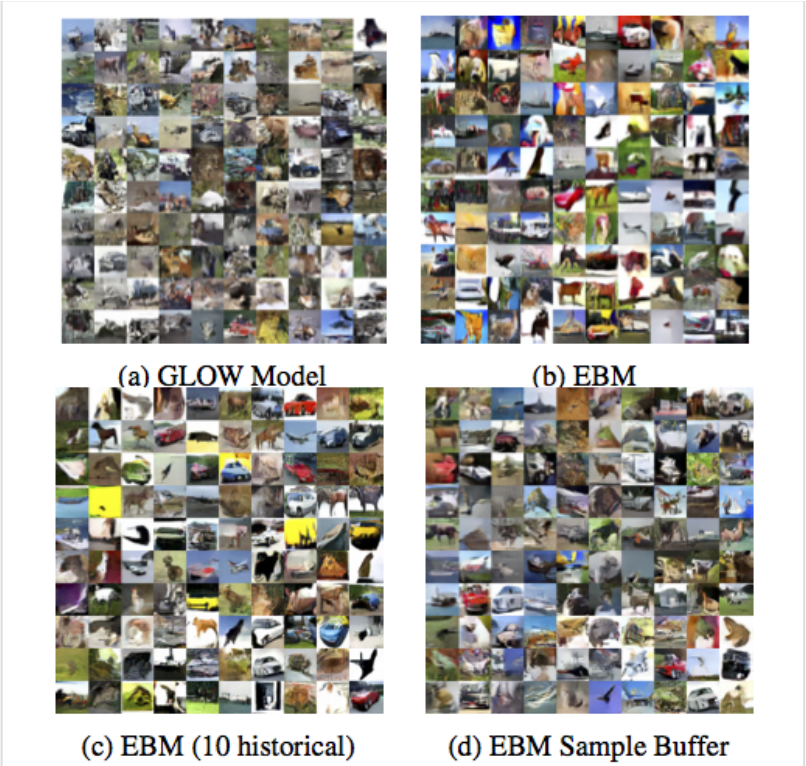

Figure 3: Comparison of image generation techniques on unconditional CIFAR-10 dataset.

在cifar10的无条件生成上的效果图