15 Feb 能量视角下的GAN模型（二）：GAN = “分析” + “采样”

By 苏剑林 | 2019-02-15 | 52518位读者

在这个系列中，我们尝试从能量的视角理解GAN。我们会发现这个视角如此美妙和直观，甚至让人拍案叫绝。

上一篇文章里，我们给出了一个直白而用力的能量图景，这个图景可以让我们轻松理解GAN的很多内容，换句话说，通俗的解释已经能让我们完成大部分的理解了，并且把最终的结论都已经写了出来。在这篇文章中，我们继续从能量的视角理解GAN，这一次，我们争取把前面简单直白的描述，用相对严密的数学语言推导一遍。

跟第一篇文章一样，对于笔者来说，这个推导过程依然直接受启发于Benjio团队的新作《Maximum Entropy Generators for Energy-Based Models》。

原作者的开源实现：https://github.com/ritheshkumar95/energy_based_generative_models

本文的大致内容如下：

1、推导了能量分布下的正负相对抗的更新公式；

2、比较了理论分析与实验采样的区别，而将两者结合便得到了GAN框架；

3、导出了生成器的补充loss，理论上可以防止mode collapse；

4、简单提及了基于能量函数的MCMC采样。

数学视角的能量

在这部分中，我们先来简单引入能量模型，并且推导了能量模型理论上的更新公式，指出它具有正相、负相对抗的特点。

能量分布模型

首先，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

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

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）

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

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”。

至于为什么选择这样的能量分布，解释有很多，既可以说是从物理角度受到启发，也可以说是从最大熵原理中受到启发，甚至你也可以简单地认为只是因为这种分布相对容易处理而已。但不可否认，这种分布很常见、很实用，我们用得非常多的softmax激活，其实也就是假设了这种分布。

现在的困难是如何求出参数θ来，而困难的来源则是配分函数(2)通常难以显式地计算出来。当然，尽管实际计算存在困难，但不妨碍我们继续把推导进行下去。

正负相的对抗

为了求出参数θ，我们先定义对数似然函数：

$$\mathrm{E}_{x\sim p(x)}\left[\log q_\theta(x)\right]$$

我们希望它越大越好，也就是希望

$$L_\theta = \mathrm{E}_{x\sim p(x)}\left[-\log q_\theta(x)\right]$$

越小越好，为此，我们对L_θ使用梯度下降。我们有

$$\begin{aligned}\nabla_\theta \log q_\theta(x) &= \nabla_\theta \log e^{-U_\theta(x)} - \nabla_\theta \log Z_\theta \\ &= -\nabla_\theta U_\theta(x) - \frac{1}{Z_\theta}\nabla_\theta Z_\theta \\ &= -\nabla_\theta U_\theta(x) - \frac{1}{Z_\theta}\nabla_\theta \int e^{-U_\theta(x)}dx \\ &= -\nabla_\theta U_\theta(x) + \frac{1}{Z_\theta}\int e^{-U_\theta(x)}\nabla_\theta U_\theta(x)dx \\ &= -\nabla_\theta U_\theta(x) + \int \frac{e^{-U_\theta(x)}}{Z_\theta}\nabla_\theta U_\theta(x)dx \\ &= -\nabla_\theta U_\theta(x) + \mathrm{E}_{x\sim q_\theta(x)}\left[\nabla_\theta U_\theta(x)\right]\end{aligned}$$

所以

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

这意味着梯度下降的更新公式是

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

注意到式(6)的特点，它是$\nabla_\theta U_\theta(x)$分别在真实分布下和拟合分布下的均值之差，这就是机器学习中著名的“正相”和“负相”的分解，式(6)体现了正负相之间的对抗，也有人将其对应为我们做梦的过程。

扬长避短 ⇒ GAN

在这部分中，我们表明“容易分析”与“容易采样”是很难兼容的，容易理论分析的模型，在实验上难以采样计算，而容易采样计算的模型，难以进行简明的理论推导。而试图将两者的优点结合起来，就得到了GAN模型。

理论分析与实验采样

事实上，式(6)和式(7)表明我们开始假设的能量分布模型的理论分析并不困难，但是落实到实验中，我们发现必须要完成从q_θ中采样：$\mathrm{E}_{x\sim q_\theta(x)}$。也就是说，给定一个具体的$U_\theta(x)$，我们要想办法从$q_\theta(x)=e^{-U_\theta(x)}/Z_\theta$中采样出一批$x$出来。

然而，就目前而言，我们对从$q_\theta(x)=e^{-U_\theta(x)}/Z_\theta$中采样并没有任何经验。对于我们来说，方便采样的是如下的过程

$$z\sim q(z),\quad x=G_\varphi(z)$$

这里的$q(z)$代表着标准正态分布。也就是说，我们可以从标准正态分布中采样出一个z出来，然后通过固定的模型G_φ变换为我们想要的x。这意味着这种分布的理论表达式是：

$$q_\varphi(x)=\int \delta\big(x-G_\varphi(z)\big)q(z)dz$$

问题是，如果用$q_\varphi(x)$代替原来的$q_\theta(x)$，那么采样是方便了，但是类似的理论推导就困难了，换句话说，我们根本推导不出类似(7)的结果来。

GAN诞生记

那么，一个异想天开的念头是：能不能把两者结合起来，在各自擅长的地方发挥各自的优势？

式(7)中的$\mathrm{E}_{x\sim q_\theta(x)}$不是难以实现吗，那我只把这部分用$\mathrm{E}_{x\sim q_\varphi(x)}$代替好了：

$$\theta\leftarrow\theta-\varepsilon\Big(\mathrm{E}_{x\sim p(x)}\big[\nabla_\theta U_\theta(x)\big]-\mathrm{E}_{x\sim q_\varphi(x)}\big[\nabla_\theta U_\theta(x)\big]\Big)$$

也就是

$$\theta\leftarrow\theta-\varepsilon\Big(\mathrm{E}_{x\sim p(x)}\big[\nabla_\theta U_\theta(x)\big]-\mathrm{E}_{x=G_\varphi(z),z\sim q(z)}\big[\nabla_\theta U_\theta(x)\big]\Big)$$

现在采样是方便了，但前提是$q_\varphi(x)$跟$q_\theta(x)$足够接近才行呀（因为$q_\theta(x)$才是标准的、正确的），所以，我们用KL散度来度量两者的差异：

$$\begin{aligned}KL\big(q_\varphi(x)\|q_\theta(x)\big)&=\int q_\varphi(x)\log\frac{q_\varphi(x)}{q_\theta(x)}dx\\&=-H_\varphi(X)+\mathrm{E}_{x\sim q_\varphi(x)}\big[U_\theta(x)\big]+\log Z_\theta\end{aligned}$$

式(11)有效的前提是$q_\varphi(x)$跟$q_\theta(x)$足够接近，也就是上式足够小，而对于固定的$q_\theta(x)$，Z_θ是一个常数，所以φ的优化目标是：

$$\varphi=\arg\min_\varphi\ -H_\varphi(X)+\mathrm{E}_{x\sim q_\varphi(x)}\big[U_\theta(x)\big]$$

这里$H_\varphi(X) = -\int q_\varphi(x)\log q_\varphi(x)dx$代表$q_\varphi(x)$的熵。$-H_\varphi(X)$希望熵越大越好，这意味着多样性；$\mathrm{E}_{x\sim q_\varphi(x)}[U_\theta(x)]$希望图片势能越小越好，这意味着真实性。

另外一方面，注意到式(11)实际上是目标

$$\theta = \arg\min_\theta \mathrm{E}_{x\sim p(x)}\left[U_\theta(x)\right] - \mathrm{E}_{x=G_\varphi(z),z\sim q(z)}\left[U_\theta(x)\right]$$

的梯度下降公式。所以我们发现，整个过程实际上就是(14)和(13)的交替梯度下降。而正如第一篇所说的，θ的这个目标可能带来数值不稳定性，基于第一篇所说的理由，真样本应该在极小值点附近，所以我们可以把梯度惩罚项补充进(14)，得到最终的流程是：

$$\begin{aligned}\theta &= \arg\min_\theta \mathrm{E}_{x\sim p(x)}\left[U_\theta(x)\right] - \mathrm{E}_{x=G_\varphi(z),z\sim q(z)}\left[U_\theta(x)\right] + \lambda\mathrm{E}_{x\sim p(x)}\left[\|\nabla_x U_\theta(x)\|^2\right]\\ \varphi &= \arg\min_\varphi -H_\varphi(X) + \mathrm{E}_{x=G_\varphi(z),z\sim q(z)}\left[U_\theta(x)\right]\end{aligned}$$

这便是基于梯度惩罚的GAN模型，我们在《能量视角下的GAN模型（一）》中已经把它“头脑风暴”出来了，而现在我们从能量模型的数学分析中把它推导出来了。

所以说，GAN实际上就是能量模型和采样模型各自扬长避短的结果。

直击$H(X)$！

现在，距离完整地实现整个模型，就差$H_\varphi(X)$了。我们已经说过

$$H_\varphi(X) = -\int q_\varphi(x)\log q_\varphi(x)dx$$

代表$q_\varphi(x)$的熵，而$q_\varphi(x)$的理论表达式是(9)，积分难以计算，所以$H_\varphi(X)$也难以计算。

打破这一困境的思路是将熵转化为互信息，然后转化为互信息的估计，其估计方式有两种：通过f散度的方式（理论上精确）估计，或者通过信息下界的方式估计。

最大熵与互信息

首先，我们可以利用$x = G_\varphi(z)$这一点：$x = G_\varphi(z)$意味着条件概率$q_\varphi(x\,|\,z) = \delta\big(x - G(z)\big)$，即一个确定性的模型，也可以理解为均值为$G(z)$、方差为0的高斯分布$\mathcal{N}(x;\, G_\varphi(z),\, 0)$。

然后我们去考虑互信息$I(X,Z)$：

$$\begin{aligned}I_\varphi(X,Z) &= \iint q_\varphi(x\,|\,z)q(z)\log\frac{q_\varphi(x\,|\,z)}{q_\varphi(x)}dxdz\\ &= \iint q_\varphi(x\,|\,z)q(z)\log q_\varphi(x\,|\,z)dxdz - \iint q_\varphi(x\,|\,z)q(z)\log q_\varphi(x)dxdz\\ &= \int q(z)\left(\int q_\varphi(x\,|\,z)\log q_\varphi(x\,|\,z)dx\right)dz + H(X)\end{aligned}$$

现在我们找出了$I_\varphi(X,Z)$和$H_\varphi(X)$的关系，它们的差是

$$\int q(z)\left(\int q_\varphi(x|z)\log q_\varphi(x|z)dx\right)dz \triangleq -H_\varphi(X|Z)$$

事实上$H_\varphi(X|Z)$称为“条件熵”。

如果我们处理的是离散型分布，那么因为$x=G_\varphi(z)$是确定性的，所以$q_\varphi(x|z)\equiv 1$，那么$H_\varphi(X|Z)$为0，即$I_\varphi(X,Z)=H_\varphi(X)$；如果是连续型分布，前面说了可以理解为方差为0的高斯分布$\mathcal{N}(x;G_\varphi(z),0)$，我们可以先考虑常数方差的情况$\mathcal{N}(x;G(z),\sigma^2)$，计算发现$H_\varphi(X|Z)\sim \log\sigma^2$是一个常数，然后$\sigma\to 0$，不过发现结果是无穷大。无穷大原则上是不能计算的，但事实上方差也不需要等于0，只要足够小，肉眼难以分辨即可。

所以，总的来说我们可以确定互信息$I_\varphi(X,Z)$与熵$H_\varphi(X)$只相差一个无关紧要的常数，所以在式(15)中，可以将$H_\varphi(X)$替换为$I_\varphi(X,Z)$：

$$\begin{aligned}\theta &= \mathop{\arg\min}_\theta \mathrm{E}_{x\sim p(x)}\left[U_\theta(x)\right] - \mathrm{E}_{x=G_\varphi(z),z\sim q(z)}\left[U_\theta(x)\right] + \lambda \mathrm{E}_{x\sim p(x)}\left[\|\nabla_x U_\theta(x)\|^2\right]\\ \varphi &= \mathop{\arg\min}_\varphi -I_\varphi(X,Z) + \mathrm{E}_{x=G_\varphi(z),z\sim q(z)}\left[U_\theta(x)\right]\end{aligned}$$

现在我们要最小化$-I_\varphi(X,Z)$，也就是最大化互信息$I_\varphi(X,Z)$。直观上这也不难理解，因为这一项是用来防止mode callopse的，而如果一旦mode callopse，那么几乎任意的z都生成同一个x，X,Z的互信息一定不会大。

但是将目标从$H_\varphi(X)$改为$I_\varphi(X,Z)$，看起来只是形式上的转换，似乎依然还没有解决问题。但很幸运的是，我们已经做过最大化互信息的研究了，方法在《深度学习的互信息：无监督提取特征》的“互信息本质”一节，也就是说，直接估算互信息已经有解决方案了，读者直接看那篇文章即可，不再重复论述。

互信息与信息下界

如果不需要精确估计互信息，那么可以使用InfoGAN中的思路，得到互信息的一个下界，然后去优化这个下界。

从互信息定义出发：

$$I_\varphi(X,Z)=\iint q_\varphi(x|z)q(z)\log\frac{q_\varphi(x|z)q(z)}{q_\varphi(x)q(z)}dxdz$$

记$q_\varphi(z|x)=q_\varphi(x|z)q(z)/q_\varphi(x)$，这代表精确的后验分布；然后对于任意近似的后验分布$p(z|x)$，我们有

$$
\begin{aligned}
I_\varphi(X,Z) &= \iint q_\varphi(x|z)q(z)\log\frac{q_\varphi(z|x)}{q(z)}dxdz \\
&= \iint q_\varphi(x|z)q(z)\log\frac{p(z|x)}{q(z)}dxdz + \iint q_\varphi(x|z)q(z)\log\frac{q_\varphi(z|x)}{p(z|x)}dxdz \\
&= \iint q_\varphi(x|z)q(z)\log\frac{p(z|x)}{q(z)}dxdz + \int q_\varphi(x)KL\Big(q_\varphi(z|x)\|p(z|x)\Big)dz \\
&\geq \iint q_\varphi(x|z)q(z)\log\frac{p(z|x)}{q(z)}dxdz \\
&= \iint q_\varphi(x|z)q(z)\log p(z|x) - \iint q_\varphi(x|z)q(z)\log q(z)dxdz \\
&\qquad\qquad \overset{\text{是一个常数}}{=} \int q(z)\log q(z)dz
\end{aligned}
$$

也就是说，互信息大于等于$\iint q_\varphi(x|z)q(z)\log p(z|x)$加上一个常数。如果最大化互信息，可以考虑最大化这个下界。由于$p(z|x)$是任意的，可以简单假设$p(z|x)=\mathcal{N}\Big(z;E(x),\sigma^2\Big)$，其中$E(x)$是一个带参数的编码器，代入计算并省去多余的常数，可以发现相当于在生成器加入一项loss：

$$\mathrm{E}_{z\sim q(z)}\big[\|z-E(G(z))\|^2\big]$$

所以，基于InfoGAN的信息下界思路，式(15)变为：

$$
\begin{aligned}
\theta &= \arg\min_\theta \mathrm{E}_{x\sim p(x)}\big[U_\theta(x)\big] - \mathrm{E}_{z\sim q(z)}\big[U_\theta(G_\varphi(z))\big] + \lambda_1\mathrm{E}_{x\sim p(x)}\big[\|\nabla_x U_\theta(x)\|^2\big] \\
\varphi, E &= \arg\min_{\varphi,E} \mathrm{E}_{z\sim q(z)}\big[U_\theta(G_\varphi(z)) + \lambda_2\|z-E(G_\varphi(z))\|^2\big]
\end{aligned}
$$

到这里，我们已经从两个角度完成了$H_\varphi(X)$的处理，从而完成了整个GAN和能量模型的推导。

MCMC提升效果

回顾开头，我们是从能量分布出发推导出了GAN模型，而能量函数$U(x)$也就是GAN模型中的判别器。既然$U(x)$具有能量函数的含义，那么训练完成后，我们可以利用能量函数的特性做更多有价值的事情，例如引入MCMC来提升效果。

MCMC的简介

其实对于MCMC，我只是略懂它的含义，并不懂它的方法和精髓，所谓“简介”，仅仅是对其概念做一些基本的介绍。MCMC是“马尔科夫链蒙特卡洛方法（Markov Chain Monte Carlo）”，在我的理解里，它大概是这么个东西：我们难以直接从某个给定的分布$q(x)$中采样出样本来，但是我们可以构造如下的随机过程：

$$x_{n+1} = f(x_n,\alpha)$$

其中α是一个便于实现的随机过程，比如从二元分布、正态分布采样等。这样一来，从某个x_0出发，得到的序列$\{x_1,x_2,\ldots,x_n,\ldots\}$是随机的。

如果进一步能证明式(24)的静态分布正好是$q(x)$，那么就意味着序列$\{x_1,x_2,\ldots,x_n,\ldots\}$正是从$q(x)$中采样出来的一批样本，这样就实现了从$q(x)$中采样了，只不过采样的结果经过了一定的顺序排列。

Langevin方程

式(24)的一个特例是Langevin方程：

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

它也称为随机微分方程，当$\varepsilon \to 0$时，它的静态分布正好是能量分布

$$p(x) = \frac{e^{-U(x)}}{Z}$$

也就是说，给定能量函数$U(x)$后，我们可以通过式(25)实现从能量分布中采样，这就是能量分布的MCMC采样的原始思想。

当然，直接从能量函数和式(25)中采样x可能不大现实，因为x维度（常见的情景下，x代表图片）过大，可控性难以保证。另一方面，式(25)最后一项是高斯噪声，所以只要$\varepsilon \neq 0$，那么结果必然是有噪声的，图片真实性也难以保证。

一个有趣的转化是：我们可以不直接考虑x的MCMC采样，而考虑z的采样。因为在前面的模型中，我们最后既得到了能量函数$U_\theta(x)$，也得到了生成模型$G_\varphi(z)$，这意味着z的能量函数为

$$U_\theta(G_\varphi(z))$$

> **注：**这个结果并不是严格成立的，只能算是一个经验公式，严格来讲只有当G的雅可比行列式为1时才成立。我也曾在github上跟作者讨论过，他也指出这没有什么严格的理论推导，只是凭直觉来的，详情可以参考：https://github.com/ritheshkumar95/energy_based_generative_models/issues/4

有了z的能量函数，我们可以通过式(25)实现z的MCMC采样：

$$z_{t+1} = z_t - \frac{1}{2}\varepsilon\nabla_z U_\theta(G_\varphi(z_t)) + \sqrt{\varepsilon}\alpha, \quad \alpha \sim \mathcal{N}(\alpha; 0, 1)$$

这样刚才说的问题全部都没有了，因为z的维度一般比x小得多，而且也不用担心$\varepsilon \neq 0$带来噪声，因为z本来就是噪声。

更好的截断技巧

到这里，如果头脑还没有混乱的读者也许会回过神来：z的分布不就是标准的正态分布吗？采样起来不是很容易吗？为啥还要折腾一套MCMC采样？

理想情况下，z的能量函数$U_\theta(G_\varphi(z))$所对应的能量分布

$$q_{\theta,\varphi}(z) = \frac{e^{-U_\theta(G_\varphi(z))}}{Z}$$

确实应该就是我们原始传递给它的标准正态分布$q(z)$。但事实上，理想和现实总有些差距的，当我们用标准正态分布去训练好一个生成模型后，最后能产生真实的样本的噪声往往会更窄一些，这就需要一些截断技巧，或者说筛选技巧。

比如，基于flow的生成模型在训练完成后，往往使用“退火”技巧，也就是在生成时将噪声的方差设置小一些，这样能生成一些更稳妥的样本，可以参考《细水长flow之NICE：流模型的基本概念与实现》。而去年发布的BigGAN，也讨论了GAN中对噪声的截断技巧。

如果我们相信我们的模型，相信能量函数$U_\theta(x)$和生成模型$G_\varphi(z)$都是有价值的，那么我们有理由相信$e^{-U_\theta(G_\varphi(z))}/Z$会是一个比标准正态分布更好的$z$的分布（能生成更真实的$x$的$z$的分布，因为它将$G_\varphi(z)$也纳入了分布的定义中），所以从$e^{-U_\theta(G_\varphi(z))}/Z$采样会优于从$q(z)$采样，也就是说MCMC采样(28)能够提升采样后的生成质量，原论文已经验证了这一点。我们可以将它理解为一种更好的截断技巧。

更高效的MALA

采样过程(28)其实依然会比较低效，原论文事实上用的是改进版本，称为MALA（Metropolis-adjusted Langevin algorithm），它在(28)的基础上进一步引入了一个筛选过程：

$$\tilde{z}_{t+1} = z_t - \frac{1}{2}\varepsilon\nabla_z U_\theta(G_\varphi(z_t)) + \sqrt{\varepsilon}\alpha, \quad \alpha \sim \mathcal{N}(\alpha; 0, 1)$$

$$z_{t+1} = \begin{cases} \tilde{z}_{t+1}, & \text{如果}\beta < \gamma \\ z_t, & \text{其他情况} \end{cases}, \quad \beta \sim U[0, 1]$$

$$\gamma = \min\left\{1, \frac{q(\tilde{z}_{t+1})q(z_t|\tilde{z}_{t+1})}{q(\tilde{z}_t)q(\tilde{z}_{t+1}|z_t)}\right\}$$

这里

$$q(z) \propto \exp\Big(-U_\theta(G_\varphi(z))\Big)$$

$$q(z'|z) \propto \exp\left(-\frac{1}{2\varepsilon}\|z' - z + \varepsilon\nabla_z U_\theta(G_\varphi(z))\|^2\right)$$

也就是说以概率γ接受$z_{t+1} = \tilde{z}_{t+1}$，以$1-\gamma$的概率保持不变。按照维基百科上的说法，这样的改进能够让采样过程更有机会采样到高概率的样本，这也就意味着能生成更多的真实样本。（笔者并不是很懂这一套理论，所以，只能照搬了~）

有力的能量视角

又是一篇公式长文，总算把能量分布下的GAN的数学推导捋清楚了，GAN是调和“理论分析”与“实验采样”矛盾的产物。总的来说，笔者觉得整个推导过程还是颇具启发性的，也能让我们明白GAN的关键之处和问题所在。

能量视角是一个偏向数学物理的视角，一旦能将机器学习和数学物理联系起来，还将可以很直接地从数学物理处获得启发，甚至使得对应的机器学习不再“黑箱”，这样的视角往往让人陶醉，给人一种有力的感觉。