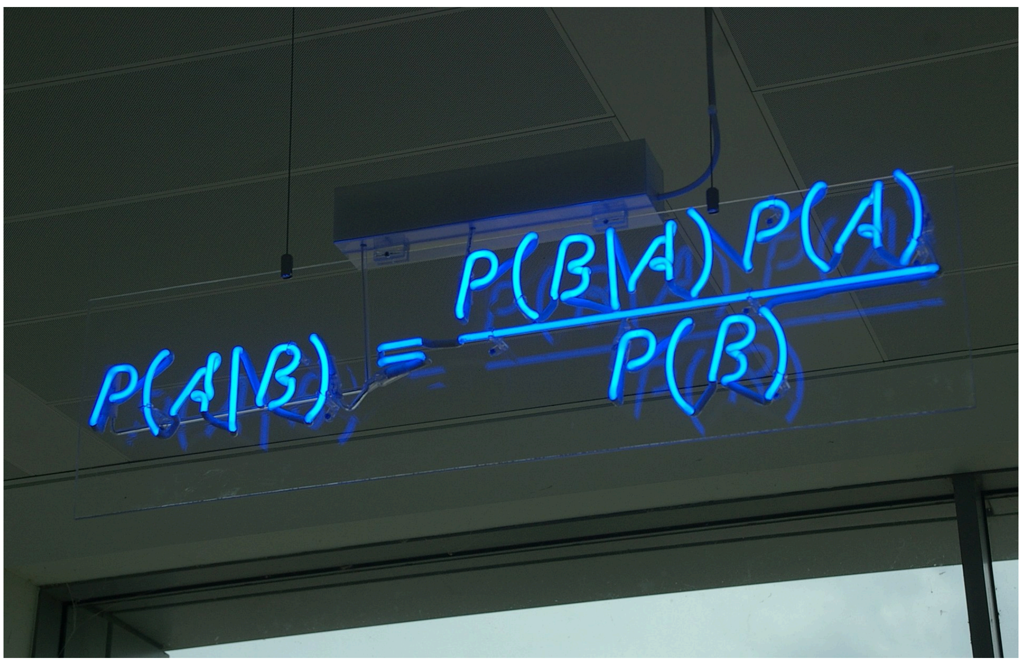
19 Jul

生成扩散模型漫谈（三）：DDPM = 贝叶斯 + 去噪

By 苏剑林 | 2022-07-19 | 135607位读者 引用

到目前为止，笔者给出了生成扩散模型DDPM的两种推导，分别是《生成扩散模型漫谈（一）：DDPM = 拆楼 + 建楼》中的通俗类比方案和《生成扩散模型漫谈（二）：DDPM = 自回归式VAE》中的变分自编码器方案。两种方案可谓各有特点，前者更为直白易懂，但无法做更多的理论延伸和定量理解，后者理论分析上更加完备一些，但稍显形式化，启发性不足。



贝叶斯定理（来自维基百科）

在这篇文章中，我们再分享DDPM的一种推导，它主要利用到了贝叶斯定理来简化计算，整个过程的“推敲”味道颇浓，很有启发性。不仅如此，它还跟我们后面将要介绍的DDIM模型有着紧密的联系。

模型绘景

再次回顾，DDPM建模的是如下变换流程：

$$\boldsymbol{x} = \boldsymbol{x}_0 \rightleftharpoons \boldsymbol{x}_1 \rightleftharpoons \boldsymbol{x}_2 \rightleftharpoons \cdots \rightleftharpoons \boldsymbol{x}_{T-1} \rightleftharpoons \boldsymbol{x}_T = \boldsymbol{z} \tag{1}$$

其中，正向就是将样本数据$\boldsymbol{x}$逐渐变为随机噪声$\boldsymbol{z}$的过程，反向就是将随机噪声$\boldsymbol{z}$逐渐变为样本数据$\boldsymbol{x}$的过程，反向过程就是我们希望得到的“生成模型”。

正向过程很简单，每一步是

$$\boldsymbol{x}_t = \alpha_t \boldsymbol{x}_{t-1} + \beta_t \boldsymbol{\varepsilon}_t,\quad \boldsymbol{\varepsilon}_t\sim\mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \tag{2}$$

或者写成$p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})=\mathcal{N}(\boldsymbol{x}_t;\alpha_t \boldsymbol{x}_{t-1},\beta_t^2 \boldsymbol{I})$。在约束$\alpha_t^2 + \beta_t^2 = 1$之下，我们有

$$\begin{aligned}
\boldsymbol{x}_t =&\, \alpha_t \boldsymbol{x}_{t-1} + \beta_t \boldsymbol{\varepsilon}_t \\
=&\, \alpha_t \big(\alpha_{t-1} \boldsymbol{x}_{t-2} + \beta_{t-1} \boldsymbol{\varepsilon}_{t-1}\big) + \beta_t \boldsymbol{\varepsilon}_t \\
=&\,\cdots\\
=&\,(\alpha_t\cdots\alpha_1) \boldsymbol{x}_0 + \underbrace{(\alpha_t\cdots\alpha_2)\beta_1 \boldsymbol{\varepsilon}_1 + (\alpha_t\cdots\alpha_3)\beta_2 \boldsymbol{\varepsilon}_2 + \cdots + \alpha_t\beta_{t-1} \boldsymbol{\varepsilon}_{t-1} + \beta_t \boldsymbol{\varepsilon}}_{\sim \mathcal{N}(\boldsymbol{0}, (1-\alpha_t^2\cdots\alpha_1^2)\boldsymbol{I})}
\end{aligned}$$

从而可以求出$p(\boldsymbol{x}_t|\boldsymbol{x}_0)=\mathcal{N}(\boldsymbol{x}_t;\bar{\alpha}_t \boldsymbol{x}_0,\bar{\beta}_t^2 \boldsymbol{I})$，其中$\bar{\alpha}_t = \alpha_1\cdots\alpha_t$，而$\bar{\beta}_t = \sqrt{1-\bar{\alpha}_t^2}$。

DDPM要做的事情，就是从上述信息中求出反向过程所需要的$p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$，这样我们就能实现从任意一个$\boldsymbol{x}_T=\boldsymbol{z}$出发，逐步采样出$\boldsymbol{x}_{T-1},\boldsymbol{x}_{T-2},\cdots,\boldsymbol{x}_1$，最后得到随机生成的样本数据$\boldsymbol{x}_0=\boldsymbol{x}$。

请贝叶斯

下面我们请出伟大的贝叶斯定理。事实上，直接根据贝叶斯定理我们有

$$p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \frac{p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})p(\boldsymbol{x}_{t-1})}{p(\boldsymbol{x}_t)} \tag{4}$$

然而，我们并不知道$p(\boldsymbol{x}_{t-1}),p(\boldsymbol{x}_t)$的表达式，所以此路不通。但我们可以退而求其次，在给定$\boldsymbol{x}_0$的条件下使用贝叶斯定理：

$$p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \frac{p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}{p(\boldsymbol{x}_t|\boldsymbol{x}_0)} \tag{5}$$

这样修改自然是因为$p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}), p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0), p(\boldsymbol{x}_t|\boldsymbol{x}_0)$都是已知的，所以上式是可计算的，代入各自的表达式得到：

$$p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) = \mathcal{N}\left(\boldsymbol{x}_{t-1}; \frac{\alpha_t\bar{\beta}_{t-1}^2}{\bar{\beta}_t^2}\boldsymbol{x}_t + \frac{\bar{\alpha}_{t-1}\beta_t^2}{\bar{\beta}_t^2}\boldsymbol{x}_0, \frac{\bar{\beta}_{t-1}^2\beta_t^2}{\bar{\beta}_t^2}\boldsymbol{I}\right) \tag{6}$$

推导：上式的推导过程并不难，就是常规的展开整理而已，当然我们也可以找点技巧加快计算。首先，代入各自的表达式，可以发现指数部分除掉$-1/2$因子外，结果是：

$$\frac{\Vert\boldsymbol{x}_t - \alpha_t\boldsymbol{x}_{t-1}\Vert^2}{\beta_t^2} + \frac{\Vert\boldsymbol{x}_{t-1} - \bar{\alpha}_{t-1}\boldsymbol{x}_0\Vert^2}{\bar{\beta}_{t-1}^2} - \frac{\Vert\boldsymbol{x}_t - \bar{\alpha}_t\boldsymbol{x}_0\Vert^2}{\bar{\beta}_t^2} \tag{7}$$

它关于$\boldsymbol{x}_{t-1}$是二次的，因此最终的分布必然也是正态分布，我们只需要求出其均值和协方差。不难看出，展开式中$\Vert\boldsymbol{x}_{t-1}\Vert^2$项的系数是

$$\frac{\alpha_t^2}{\beta_t^2} + \frac{1}{\bar{\beta}_{t-1}^2} = \frac{\alpha_t^2\bar{\beta}_{t-1}^2 + \beta_t^2}{\bar{\beta}_{t-1}^2\beta_t^2} = \frac{\alpha_t^2(1-\bar{\alpha}_{t-1}^2) + \beta_t^2}{\bar{\beta}_{t-1}^2\beta_t^2} = \frac{1-\bar{\alpha}_t^2}{\bar{\beta}_{t-1}^2\beta_t^2} = \frac{\bar{\beta}_t^2}{\bar{\beta}_{t-1}^2\beta_t^2} \tag{8}$$

所以整理好的结果必然是$\frac{\bar{\beta}_t^2}{\bar{\beta}_{t-1}^2\beta_t^2}\Vert\boldsymbol{x}_{t-1} - \tilde{\boldsymbol{\mu}}(\boldsymbol{x}_t, \boldsymbol{x}_0)\Vert^2$的形式，这意味着协方差矩阵是$\frac{\bar{\beta}_{t-1}^2\beta_t^2}{\bar{\beta}_t^2}\boldsymbol{I}$。另一边，把一次项系数拿出来是$-2\left(\frac{\alpha_t}{\beta_t^2}\boldsymbol{x}_t + \frac{\bar{\alpha}_{t-1}}{\bar{\beta}_{t-1}^2}\boldsymbol{x}_0\right)$，除以$\frac{-2\bar{\beta}_t^2}{\bar{\beta}_{t-1}^2\beta_t^2}$后便可以得到

$$\tilde{\boldsymbol{\mu}}(\boldsymbol{x}_t, \boldsymbol{x}_0) = \frac{\alpha_t\bar{\beta}_{t-1}^2}{\bar{\beta}_t^2}\boldsymbol{x}_t + \frac{\bar{\alpha}_{t-1}\beta_t^2}{\bar{\beta}_t^2}\boldsymbol{x}_0 \tag{9}$$

这就得到了$p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$的所有信息了，结果正是式(6)。

去噪过程

现在我们得到了$p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$，它有显式的解，但并非我们想要的最终答案，因为我们只想通过$\boldsymbol{x}_t$来预测$\boldsymbol{x}_{t-1}$，而不能依赖$\boldsymbol{x}_0$，$\boldsymbol{x}_0$是我们最终想要生成的结果。接下来，

一个“异想天开”的想法是

> 如果我们能够通过$\boldsymbol{x}_t$来预测$\boldsymbol{x}_0$，那么不就可以消去$p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$中的$\boldsymbol{x}_0$，使得它只依赖于$\boldsymbol{x}_t$了吗？

说干就干，我们用$\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t)$来预估$\boldsymbol{x}_0$，损失函数为$\Vert\boldsymbol{x}_0 - \bar{\boldsymbol{\mu}}(\boldsymbol{x}_t)\Vert^2$。训练完成后，我们就认为

$$p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) \approx p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0=\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t)) = \mathcal{N}\left(\boldsymbol{x}_{t-1}; \frac{\alpha_t\bar{\beta}_{t-1}^2}{\bar{\beta}_t^2}\boldsymbol{x}_t + \frac{\bar{\alpha}_{t-1}\beta_t^2}{\bar{\beta}_t^2}\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t), \frac{\bar{\beta}_{t-1}^2}{\bar{\beta}_t^2}\right.$$

在$\Vert\boldsymbol{x}_0 - \bar{\boldsymbol{\mu}}(\boldsymbol{x}_t)\Vert^2$中，$\boldsymbol{x}_0$代表原始数据，$\boldsymbol{x}_t$代表带噪数据，所以这实际上在训练一个去噪模型，这也就是DDPM的第一个“D”的含义（Denoising）。

具体来说，$p(\boldsymbol{x}_t|\boldsymbol{x}_0)=\mathcal{N}(\boldsymbol{x}_t;\bar{\alpha}_t\boldsymbol{x}_0,\bar{\beta}_t^2\boldsymbol{I})$意味着$\boldsymbol{x}_t = \bar{\alpha}_t\boldsymbol{x}_0 + \bar{\beta}_t\boldsymbol{\varepsilon},\boldsymbol{\varepsilon}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})$，或者写成$\boldsymbol{x}_0 = \frac{1}{\bar{\alpha}_t}\left(\boldsymbol{x}_t - \bar{\beta}_t\boldsymbol{\varepsilon}\right)$，这启发我们将$\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t)$参数化为

$$\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t) = \frac{1}{\bar{\alpha}_t}\left(\boldsymbol{x}_t - \bar{\beta}_t\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)\right) \tag{11}$$

此时损失函数变为

$$\Vert\boldsymbol{x}_0 - \bar{\boldsymbol{\mu}}(\boldsymbol{x}_t)\Vert^2 = \frac{\bar{\beta}_t^2}{\bar{\alpha}_t^2}\left\Vert\boldsymbol{\varepsilon} - \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\bar{\alpha}_t\boldsymbol{x}_0 + \bar{\beta}_t\boldsymbol{\varepsilon}, t)\right\Vert^2 \tag{12}$$

省去前面的系数，就得到DDPM原论文所用的损失函数了。可以发现，本文是直接得出了从$\boldsymbol{x}_t$到$\boldsymbol{x}_0$的去噪过程，而不是像之前两篇文章那样，通过$\boldsymbol{x}_t$到$\boldsymbol{x}_{t-1}$的去噪过程再加上积分变换来推导，相比之下本文的推导可谓更加一步到位了。

另一边，我们将式(11)代入到式(10)中，化简得到

$$p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) \approx p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0=\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t)) = \mathcal{N}\left(\boldsymbol{x}_{t-1}; \frac{1}{\alpha_t}\left(\boldsymbol{x}_t - \frac{\beta_t^2}{\bar{\beta}_t}\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{x}_t, t)\right), \frac{\bar{\beta}_{t-1}^2}{\bar{\beta}_t^2}\right.$$

这就是反向的采样过程所用的分布，连同采样过程所用的方差也一并确定下来了。至此，DDPM推导完毕～（**提示：**出于推导的流畅性考虑，本文的$\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$跟前两篇介绍不一样，反而跟DDPM原论文一致。）

推导：将式(11)代入到式(10)的主要化简难度就是计算

$$\frac{\alpha_t\bar{\beta}_{t-1}^2}{\bar{\beta}_t^2}+\frac{\bar{\alpha}_{t-1}\beta_t^2}{\bar{\alpha}_t\bar{\beta}_t^2}=\frac{\alpha_t\bar{\beta}_{t-1}^2+\beta_t^2/\alpha_t}{\bar{\beta}_t^2}=\frac{\alpha_t^2(1-\bar{\alpha}_{t-1}^2)+\beta_t^2}{\alpha_t\bar{\beta}_t^2}=\frac{1-\bar{\alpha}_t^2}{\alpha_t\bar{\beta}_t^2}=\frac{1}{\alpha_t}$$

预估修正

不知道读者有没有留意到一个有趣的地方：我们要做的事情，就是想将$\boldsymbol{x}_T$慢慢地变为$\boldsymbol{x}_0$，而我们在借用$p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)$近似$p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$时，却包含了“用$\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t)$来预估$\boldsymbol{x}_0$”这一步，要是能预估准的话，那就直接一步到位了，还需要逐步采样吗?

真实情况是，“用$\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t)$来预估$\boldsymbol{x}_0$”当然不会太准的，至少开始的相当多步内不会太准。它仅仅起到了一个前瞻性的预估作用，然后我们只用$p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$来推进一小步，这就是很多数值算法中的“预估-修正”思想，即我们用一个粗糙的解往前推很多步，然后利用这个粗糙的结果将最终结果推进一小步，以此来逐步获得更为精细的解。

由此我们还可以联想到Hinton三年前提出的《Lookahead Optimizer: k steps forward, 1 step back》，它同样也包含了预估（k steps forward）和修正（1 step back）两部分，原论文将其诠释为“快（Fast）-慢（Slow）”权重的相互结合，快权重就是预估得到的结果，慢权重则是基于预估所做的修正结果。如果愿意，我们也可以用同样的方式去诠释DDPM的“预估-修正”过程～

遗留问题

最后，在使用贝叶斯定理一节中，我们说式(4)没法直接用的原因是$p(\boldsymbol{x}_{t-1})$和$p(\boldsymbol{x}_t)$均不知道。因为根据定义，我们有

$$p(\boldsymbol{x}_t) = \int p(\boldsymbol{x}_t|\boldsymbol{x}_0)\tilde{p}(\boldsymbol{x}_0)d\boldsymbol{x}_0 \tag{15}$$

其中$p(\boldsymbol{x}_t|\boldsymbol{x}_0)$是知道的，而数据分布$\tilde{p}(\boldsymbol{x}_0)$无法提前预知，所以不能进行计算。不过，有两个特殊的例子，是可以直接将两者算出来的，这里我们也补充计算一下，其结果也正好是上一篇文章遗留的方差选取问题的答案。

第一个例子是整个数据集只有一个样本，不失一般性，假设该样本为$\boldsymbol{0}$，此时$\tilde{p}(\boldsymbol{x}_0)$为狄拉克分布$\delta(\boldsymbol{x}_0)$，可以直接算出$p(\boldsymbol{x}_t) = p(\boldsymbol{x}_t|\boldsymbol{0})$。继而代入式(4)，可以发现结果正好是$p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$取$\boldsymbol{x}_0 = \boldsymbol{0}$的特例，即

$$p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0 = \boldsymbol{0}) = \mathcal{N}\left(\boldsymbol{x}_{t-1}; \frac{\alpha_t\bar{\beta}_{t-1}^2}{\bar{\beta}_t^2}\boldsymbol{x}_t, \frac{\bar{\beta}_{t-1}^2\beta_t^2}{\bar{\beta}_t^2}\boldsymbol{I}\right) \tag{16}$$

我们主要关心其方差为$\frac{\bar{\beta}_{t-1}^2\beta_t^2}{\bar{\beta}_t^2}$，这便是采样方差的选择之一。

第二个例子是数据集服从标准正态分布，即$\tilde{p}(\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_0; \boldsymbol{0}, \boldsymbol{I})$。前面我们说了$p(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \bar{\alpha}_t\boldsymbol{x}_0, \bar{\beta}_t^2\boldsymbol{I})$意味着$\boldsymbol{x}_t = \bar{\alpha}_t\boldsymbol{x}_0 + \bar{\beta}_t\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$，而此时根据假设还有$\boldsymbol{x}_0 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$，所以由正态分布的叠加性，$\boldsymbol{x}_t$正好也服从标准正态分布。将标准正态分布的概率密度代入式(4)后，结果的指数部分除掉$-1/2$因子外，结果是：

$$\frac{\|\boldsymbol{x}_t - \alpha_t\boldsymbol{x}_{t-1}\|^2}{\beta_t^2} + \|\boldsymbol{x}_{t-1}\|^2 - \|\boldsymbol{x}_t\|^2 \tag{17}$$

跟推导$p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0)$的过程类似，可以得到上述指数对应于

$$p(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \mathcal{N}\left(\boldsymbol{x}_{t-1}; \alpha_t\boldsymbol{x}_t, \beta_t^2\boldsymbol{I}\right) \tag{18}$$

我们同样主要关心其方差为β_t^2，这便是采样方差的另一个选择。

文章小结

本文分享了DDPM的一种颇有“推敲”味道的推导，它借助贝叶斯定理来直接推导反向的生成过程，相比之前的“拆楼-建楼”类比和变分推断理解更加一步到位。同时，它也更具启发性，跟接下来要介绍的DDIM有很密切的联系。

如果您需要引用本文，请参考：

苏剑林. (Jul. 19, 2022). 《生成扩散模型漫谈（三）：DDPM = 贝叶斯 + 去噪 》[Blog post]. Retrieved from https://spaces.ac.cn/archives/9164

```
@online{kexuefm-9164,
        title={生成扩散模型漫谈（三）：DDPM = 贝叶斯 + 去噪},
        author={苏剑林},
        year={2022},
        month={Jul},
        url={\url{https://spaces.ac.cn/archives/9164}},
}
```