生成扩散模型漫谈（二十五）：基于恒等式的蒸馏（上）

1 May By 苏剑林 | 2024-05-01 | 45110位读者 引用

今天我们分享一下论文《Score identity Distillation: Exponentially Fast Distillation of Pretrained Diffusion Models for One-Step Generation》，顾名思义，这是一篇探讨如何更快更好地蒸馏扩散模型的新论文。

即便没有做过蒸馏，大家应该也能猜到蒸馏的常规步骤：随机采样大量输入，然后用扩散模型生成相应结果作为输出，用这些输入输出作为训练数据对，来监督训练一个新模型。然而，众所周知作为教师的原始扩散模型通常需要多步（比如1000步）迭代才能生成高质量输出，所以且不论中间训练细节如何，该方案的一个显著缺点是生成训练数据太费时费力。此外，蒸馏之后的学生模型通常或多或少都有效果损失。

有没有方法能一次性解决这两个缺点呢？这就是上述论文试图要解决的问题。

重现江湖

论文将所提方案称为“Score identity Distillation（SiD）”，该名字取自它基于几个恒等式（Identity）来设计和推导了整个框架，取这个略显随意的名字大体是想突出恒等式变换在SiD中的关键作用，这确实是SiD的核心贡献。

至于SiD的训练思想，其实跟之前在《从去噪自编码器到生成模型》介绍过的论文《Learning Generative Models using Denoising Density Estimators》（简称“DDE”）几乎一模一样，甚至最终形式也有五六分相似。只不过当时扩散模型还未露头角，所以DDE是将其作为一种新的生成模型提出的，在当时反而显得非常小众。而在扩散模型流行的今天，它可以重新表述为一种扩散模型的蒸馏方法，因为它需要一个训练好的去噪自编码器——这正好是扩散模型的核心。

接下来笔者用自己的思路去介绍SiD。假设我们有一个在目标数据集训练好的教师扩散模型$\boldsymbol{\epsilon}_{\varphi^*}(\boldsymbol{x}_t, t)$，它需要多步采样才能生成高质量图片，我们的目标则是要训练一个单步采样的学生模型$\boldsymbol{x} = \boldsymbol{g}_{\boldsymbol{\theta}}(\boldsymbol{z})$，也就是一个类似GAN的生成器，输入指定噪声$\boldsymbol{z}$就可以

直接生成符合要求的图像。如果我们有很多的$(\boldsymbol{z}, \boldsymbol{x})$对，那么直接监督训练就可以了（当然损失函数和其他细节还需要进一步确定，读者可以自行参考相关工作），但如果没有呢？肯定不是不能训，因为就算没有$\boldsymbol{\epsilon}_{\varphi^*}(\boldsymbol{x}_t, t)$也能训，比如GAN，所以关键是怎么借助已经训练好的扩散模型提供更好的信号。

SiD及前作DDE使用了一个看上去很绕但是也很聪明的思路：

> 如果$\boldsymbol{g}_{\boldsymbol{\theta}}(\boldsymbol{z})$产生的数据分布跟目标分布很相似，那么拿$\boldsymbol{g}_{\boldsymbol{\theta}}(\boldsymbol{z})$生成的数据集去训练一个扩散模型$\boldsymbol{\epsilon}_{\psi^*}(\boldsymbol{x}_t, t)$的话，它也应该跟$\boldsymbol{\epsilon}_{\varphi^*}(\boldsymbol{x}_t, t)$很相似？

初级形式

这个思路的聪明之处在于，它绕开了对教师模型生成样本的需求，也不需要训练教师模型的真实样本，因为“拿$\boldsymbol{g}_{\boldsymbol{\theta}}(\boldsymbol{z})$生成的数据集去训练一个扩散模型”只需要学生模型$\boldsymbol{g}_{\boldsymbol{\theta}}(\boldsymbol{z})$生成的数据（简称“学生数据”），而$\boldsymbol{g}_{\boldsymbol{\theta}}(\boldsymbol{z})$是一个单步模型，用它来生成数据时间上比较友好。

当然，这还只是思路，将其转换为实际可行的训练方案还有一段路要走。首先回顾一下扩散模型，我们采用《生成扩散模型漫谈（三）：DDPM = 贝叶斯 + 去噪》的形式，我们使用如下方式对输入$\boldsymbol{x}_0$进行加噪：

$$\boldsymbol{x}_t = \bar{\alpha}_t \boldsymbol{x}_0 + \bar{\beta}_t \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}) \tag{1}$$

换言之$p(\boldsymbol{x}_t|\boldsymbol{x}_0) = \mathcal{N}(\boldsymbol{x}_t; \bar{\alpha}_t \boldsymbol{x}_0, \bar{\beta}_t^2 \boldsymbol{I})$。训练$\boldsymbol{\epsilon}_{\varphi^*}(\boldsymbol{x}_t, t)$的方式则是去噪：

$$\varphi^* = \mathop{\text{argmin}}_{\varphi} \mathbb{E}_{\boldsymbol{x}_0 \sim \tilde{p}(\boldsymbol{x}_0), \boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})} \left[\left\| \boldsymbol{\epsilon}_{\varphi}(\bar{\alpha}_t \boldsymbol{x}_0 + \bar{\beta}_t \boldsymbol{\varepsilon}, t) - \boldsymbol{\varepsilon} \right\|^2\right] \tag{2}$$

这里的$\tilde{p}(\boldsymbol{x}_0)$就是教师模型的训练数据。同样地，如果我们想用$\boldsymbol{g}_{\boldsymbol{\theta}}(\boldsymbol{z})$的学生数据一个扩散模型，那么训练目标是

$$\begin{aligned}\boldsymbol{\psi}^* &= \mathop{\text{argmin}}_{\boldsymbol{\psi}} \mathbb{E}_{\boldsymbol{x}_0^{(g)}\sim p_{\boldsymbol{\theta}}(\boldsymbol{x}_0^{(g)}),\boldsymbol{\varepsilon}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})}\left[\left\|\boldsymbol{\epsilon}_{\boldsymbol{\psi}}(\boldsymbol{x}_t^{(g)},t)-\boldsymbol{\varepsilon}\right\|^2\right] \\ &= \mathop{\text{argmin}}_{\boldsymbol{\psi}} \mathbb{E}_{\boldsymbol{z},\boldsymbol{\varepsilon}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})}\left[\left\|\boldsymbol{\epsilon}_{\boldsymbol{\psi}}(\boldsymbol{x}_t^{(g)},t)-\boldsymbol{\varepsilon}\right\|^2\right]\end{aligned} \tag{3}$$

这里$\boldsymbol{x}_t^{(g)} = \bar{\alpha}_t \boldsymbol{x}_0^{(g)} + \bar{\beta}_t\boldsymbol{\varepsilon} = \bar{\alpha}_t \boldsymbol{g}_{\boldsymbol{\theta}}(\boldsymbol{z}) + \bar{\beta}_t\boldsymbol{\varepsilon}$，是由学生数据加噪后的样本，学生数据的分布记为$p_{\boldsymbol{\theta}}(\boldsymbol{x}_0^{(g)})$；第二个等号用到了“$\boldsymbol{x}_0^{(g)}$直接由$\boldsymbol{z}$决定”的事实，所以对$\boldsymbol{x}_0^{(g)}$的期望等价于对$\boldsymbol{z}$的期望。现在我们有两个扩散模型，它们之间的差异一定程度上衡量了教师模型和学生模型生成的数据分布差异，所以一个直观的想法是通过最小化它们之间的差异，来学习学生模型：

$$\boldsymbol{\theta}^* = \mathop{\text{argmin}}_{\boldsymbol{\theta}} \underbrace{\mathbb{E}_{\boldsymbol{z},\boldsymbol{\varepsilon}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})}\left[\left\|\boldsymbol{\epsilon}_{\boldsymbol{\varphi}^*}(\boldsymbol{x}_t^{(g)},t)-\boldsymbol{\epsilon}_{\boldsymbol{\psi}^*}(\boldsymbol{x}_t^{(g)},t)\right\|^2\right]}_{\mathcal{L}_1} \tag{4}$$

注意式(3)的优化依赖于$\boldsymbol{\theta}$，所以当$\boldsymbol{\theta}$通过式(4)发生改变时，$\boldsymbol{\psi}^*$的值也随之改变，因此式(3)和式(4)实际上需要交替优化，类似GAN一样。

点睛之笔

谈到GAN，有读者可能会“闻之色变”，因为它是出了名的容易训崩。很遗憾，上述提出的式(3)和式(4)交替训练的方案同样有这个问题。首先它理论上是没有问题的，问题出现在理论与实践之间的gap，主要体现在两点：

> 1、理论上要求先求出式(3)的最优解，然后才去优化式(4)，但实际上从训练成本考虑，我们并没有将它训练到最优就去优化式(4)了；
>
> 2、理论上$\boldsymbol{\psi}^*$随$\boldsymbol{\theta}$而变，即应该写成$\boldsymbol{\psi}^*(\boldsymbol{\theta})$，从而在优化式(4)时应该多出一项$\boldsymbol{\psi}^*(\boldsymbol{\theta})$对$\boldsymbol{\theta}$的梯度，但实际上在优化式(4)时我们都只当$\boldsymbol{\psi}^*$是常数。

这两个问题非常本质，它们也是GAN训练不稳定的根本原因，此前论文《Revisiting GANs by Best-Response Constraint: Perspective, Methodology, and Application》也特意从第2点出发改进了GAN的训练。看上去，这两个问题哪一个都无法解决，尤其是第

1个，我们几乎不可能总是将ψ求到最优，这在成本上是绝对无法接受的，至于第2个，在交替训练场景下我们也没什么好办法获得$\psi^*(\boldsymbol{\theta})$的任何有效信息，从而更加不可能获得它关于$\boldsymbol{\theta}$的梯度。

幸运的是，对于上述扩散模型的蒸馏问题，SiD提出了一个有效缓解这两个问题的方案。SiD的想法可谓非常“朴素”：**既然ψ^*取近似值和ψ^*当成常数都没法避免，那么唯一的办法就是通过恒等变换，尽量消除优化目标(4)对ψ^*的依赖了。**只要式(4)对ψ^*的依赖足够弱，那么上述两个问题带来的负面影响也能足够弱了。

这就是SiD的核心贡献，也是让人拍案叫绝的“点睛之笔”。

恒等变换

接下来我们具体来看做了什么恒等变换。我们先来看式(2)，它的优化目标可以等价地改写成

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}_0\sim\tilde{p}(\boldsymbol{x}_0),\boldsymbol{\varepsilon}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})}\left[\left\|\boldsymbol{\epsilon}_{\boldsymbol{\varphi}}(\bar{\alpha}_t\boldsymbol{x}_0+\bar{\beta}_t\boldsymbol{\varepsilon},t)-\boldsymbol{\varepsilon}\right\|^2\right]\\
=&\mathbb{E}_{\boldsymbol{x}_0\sim\tilde{p}(\boldsymbol{x}_0),\boldsymbol{x}_t\sim p(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[\left\|\boldsymbol{\epsilon}_{\boldsymbol{\varphi}}(\boldsymbol{x}_t,t)-\frac{\boldsymbol{x}_t-\bar{\alpha}_t\boldsymbol{x}_0}{\bar{\beta}_t}\right\|^2\right]\\
=&\mathbb{E}_{\boldsymbol{x}_0\sim\tilde{p}(\boldsymbol{x}_0),\boldsymbol{x}_t\sim p(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[\left\|\boldsymbol{\epsilon}_{\boldsymbol{\varphi}}(\boldsymbol{x}_t,t)+\bar{\beta}_t\nabla_{\boldsymbol{x}_t}\log p(\boldsymbol{x}_t|\boldsymbol{x}_0)\right\|^2\right]
\end{aligned}\tag{5}$$

根据《生成扩散模型漫谈（五）：一般框架之SDE篇》的得分匹配相关结果，上述目标的最优解是$\boldsymbol{\epsilon}_{\boldsymbol{\varphi}^*}(\boldsymbol{x}_t,t)=-\bar{\beta}_t\nabla_{\boldsymbol{x}_t}\log p(\boldsymbol{x}_t)$，同理式$(3)$的最优解是$\boldsymbol{\epsilon}_{\boldsymbol{\psi}^*}(\boldsymbol{x}_t^{(g)},t)=-\bar{\beta}_t\nabla_{\boldsymbol{x}_t^{(g)}}\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_t^{(g)})$。此时式$(4)$的目标函数可以等价地改写成

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{z},\boldsymbol{\varepsilon}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})}\left[\left\|\boldsymbol{\epsilon}_{\boldsymbol{\varphi}^*}(\boldsymbol{x}_t^{(g)},t)-\boldsymbol{\epsilon}_{\boldsymbol{\psi}^*}(\boldsymbol{x}_t^{(g)},t)\right\|^2\right]\\
=&\mathbb{E}_{\boldsymbol{z},\boldsymbol{\varepsilon}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})}\left[\left\langle\boldsymbol{\epsilon}_{\boldsymbol{\varphi}^*}(\boldsymbol{x}_t^{(g)},t)-\boldsymbol{\epsilon}_{\boldsymbol{\psi}^*}(\boldsymbol{x}_t^{(g)},t),\boldsymbol{\epsilon}_{\boldsymbol{\varphi}^*}(\boldsymbol{x}_t^{(g)},t)+\bar{\beta}_t\nabla_{\boldsymbol{x}_t^{(g)}}\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_t^{(g)})\right\rangle\right]\\
=&\mathbb{E}_{\boldsymbol{z},\boldsymbol{\varepsilon}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})}\left[\left\langle\boldsymbol{\epsilon}_{\boldsymbol{\varphi}^*}(\boldsymbol{x}_t^{(g)},t)-\boldsymbol{\epsilon}_{\boldsymbol{\psi}^*}(\boldsymbol{x}_t^{(g)},t),\boldsymbol{\epsilon}_{\boldsymbol{\varphi}^*}(\boldsymbol{x}_t^{(g)},t)\right\rangle\right]\\
&\qquad+\mathbb{E}_{\boldsymbol{x}_t^{(g)}\sim p_{\boldsymbol{\theta}}(\boldsymbol{x}_t^{(g)})}\left[\left\langle\boldsymbol{\epsilon}_{\boldsymbol{\varphi}^*}(\boldsymbol{x}_t^{(g)},t)-\boldsymbol{\epsilon}_{\boldsymbol{\psi}^*}(\boldsymbol{x}_t^{(g)},t),\bar{\beta}_t\nabla_{\boldsymbol{x}_t^{(g)}}\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_t^{(g)})\right\rangle\right]
\end{aligned}$$

接下来要用到在《生成扩散模型漫谈（十八）：得分匹配 = 条件得分匹配》证明过的一个恒等式，来化简上式的红色部分：

$$\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t) = \mathbb{E}_{\boldsymbol{x}_0\sim p(\boldsymbol{x}_0|\boldsymbol{x}_t)}\left[\nabla_{\boldsymbol{x}_t} \log p(\boldsymbol{x}_t|\boldsymbol{x}_0)\right] \tag{7}$$

这是由概率密度定义以及贝叶斯公式推出的恒等式，不依赖于$p(\boldsymbol{x}_t), p(\boldsymbol{x}_t|\boldsymbol{x}_0), p(\boldsymbol{x}_0|\boldsymbol{x}_t)$的形式。将该恒等式代入到红色部分，我们有

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{x}_t^{(g)}\sim p_{\boldsymbol{\theta}}(\boldsymbol{x}_t^{(g)})}\left[\left\langle \boldsymbol{\epsilon}_{\boldsymbol{\varphi}^*}(\boldsymbol{x}_t^{(g)}, t) - \boldsymbol{\epsilon}_{\boldsymbol{\psi}^*}(\boldsymbol{x}_t^{(g)}, t), \bar{\beta}_t \nabla_{\boldsymbol{x}_t^{(g)}} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}_t^{(g)})\right\rangle\right] \\
=&\mathbb{E}_{\boldsymbol{x}_t^{(g)}\sim p_{\boldsymbol{\theta}}(\boldsymbol{x}_t^{(g)}),\boldsymbol{x}_0^{(g)}\sim p_{\boldsymbol{\theta}}(\boldsymbol{x}_0^{(g)}|\boldsymbol{x}_t^{(g)})}\left[\left\langle \boldsymbol{\epsilon}_{\boldsymbol{\varphi}^*}(\boldsymbol{x}_t^{(g)}, t) - \boldsymbol{\epsilon}_{\boldsymbol{\psi}^*}(\boldsymbol{x}_t^{(g)}, t), \bar{\beta}_t \nabla_{\boldsymbol{x}_t^{(g)}} \log p(\boldsymbol{x}_t^{(g)}|\boldsymbol{x}_0^{(g)})\right\rangle\right] \\
=&-\mathbb{E}_{\boldsymbol{x}_0^{(g)}\sim p_{\boldsymbol{\theta}}(\boldsymbol{x}_0^{(g)}),\boldsymbol{x}_t^{(g)}\sim p_{\boldsymbol{\theta}}(\boldsymbol{x}_t^{(g)}|\boldsymbol{x}_0^{(g)})}\left[\left\langle \boldsymbol{\epsilon}_{\boldsymbol{\varphi}^*}(\boldsymbol{x}_t^{(g)}, t) - \boldsymbol{\epsilon}_{\boldsymbol{\psi}^*}(\boldsymbol{x}_t^{(g)}, t), \frac{\boldsymbol{x}_t - \bar{\alpha}_t \boldsymbol{x}_0}{\bar{\beta}_t}\right\rangle\right] \\
=&-\mathbb{E}_{\boldsymbol{z},\boldsymbol{\varepsilon}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})}\left[\left\langle \boldsymbol{\epsilon}_{\boldsymbol{\varphi}^*}(\boldsymbol{x}_t^{(g)}, t) - \boldsymbol{\epsilon}_{\boldsymbol{\psi}^*}(\boldsymbol{x}_t^{(g)}, t), \boldsymbol{\varepsilon}\right\rangle\right]
\end{aligned}$$

跟绿色部分合并，就得到学生模型新的损失函数

$$\mathcal{L}_2 = \mathbb{E}_{\boldsymbol{z},\boldsymbol{\varepsilon}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})}\left[\left\langle \boldsymbol{\epsilon}_{\boldsymbol{\varphi}^*}(\boldsymbol{x}_t^{(g)}, t) - \boldsymbol{\epsilon}_{\boldsymbol{\psi}^*}(\boldsymbol{x}_t^{(g)}, t), \boldsymbol{\epsilon}_{\boldsymbol{\varphi}^*}(\boldsymbol{x}_t^{(g)}, t) - \boldsymbol{\varepsilon}\right\rangle\right] \tag{9}$$

这就是SiD的核心结果，原论文的实验结果显示它能够高效地实现蒸馏，而式(4)则没有训练出有意义的结果。

相比式(4)，上式(9)出现$\boldsymbol{\psi}^*$的次数显然更少，也就是对$\boldsymbol{\psi}^*$的依赖更弱。此外，上式是基于最优解$\boldsymbol{\epsilon}_{\boldsymbol{\psi}^*}(\boldsymbol{x}_t^{(g)}, t) = -\bar{\beta}_t \nabla_{\boldsymbol{x}_t^{(g)}} \log p_{\boldsymbol{\theta}}(\boldsymbol{x}_t^{(g)})$恒等变换而来的，也就是说相当于（部分地）预先窥见了$\boldsymbol{\psi}^*$的精确值，这也是它更优越的原因之一

其他细节

到目前为止，本文的推导基本上是原论文推导的重复，但出了个别记号上的不一致外，还有一些细节上的不同，下面简单澄清一下，以免读者混淆。

首先，论文的推导默认了$\bar{\alpha}_t = 1$，这是沿用了《Elucidating the Design Space of Diffusion-Based Generative Models》一文的设置。然而尽管$\bar{\alpha}_t = 1$很有代表性，并且能简化形式，但并不能很好地覆盖所有扩散模型类型，所以本文的推导保留了$\bar{\alpha}_t$。其次，论文的结果是以$\bar{\boldsymbol{\mu}}(\boldsymbol{x}_t) = \frac{\boldsymbol{x}_t - \bar{\beta}_t \boldsymbol{\epsilon}(\boldsymbol{x}_t, t)}{\bar{\alpha}_t}$为标准给出的，这显然跟扩散模型常见的以$\boldsymbol{\epsilon}(\boldsymbol{x}_t, t)$为准不符，笔者暂时没有领悟到原论文的表述方式的优越所在。

最后，原论文发现损失函数$\mathcal{L}_1$即(4)实在太不稳定，往往对效果还起到负面作用，所以SiD最终取了式(4)的相反数作为额外的损失函数，加权到改进的损失函数(9)上，即最终损失为$\mathcal{L}_2 - \lambda \mathcal{L}_1$（注：原论文中的权重记号是$\alpha$，但本文$\alpha$已用来表示noise schedule，所以改用$\lambda$），这在个别情形还能取得更优的蒸馏效果。至于具体实验细节和数据，读者自行翻阅原论文就好。

相比其他蒸馏方法，SiD的缺点是对显存的需求比较大，因为它同时要维护三个模型$\boldsymbol{\epsilon}_{\boldsymbol{\varphi}}(\boldsymbol{x}_t, t)$、$\boldsymbol{\epsilon}_{\boldsymbol{\psi}}(\boldsymbol{x}_t, t)$和$\boldsymbol{g}_{\boldsymbol{\theta}}(\boldsymbol{z})$，它们具有相同的体量，虽然并非同时进行反向传播，但叠加起来也使得总显存量翻了一倍左右。针对这个问题，SiD在正文末尾提出，未来可以尝试对预训练的模型加LoRA来作为额外引入的两个模型，以进一步节省显存需求。

延伸思考

笔者相信，对于一开始的“初级形式”，即式(3)和式(4)的交替优化，那么不少理论基础比较扎实并且深入思考过的读者都有机会想到，尤其是已经有DDE“珠玉在前”，推出它似乎并不是那么难预估的事情。但SiD的精彩之处是并没有止步于此，而是提出了后面的恒等变换，使得训练更加稳定高效，这体现了作者对扩散模型和优化理论非常深刻的理解。

同时，SiD也留下了不少值得进一步思考和探索的问题。比如，学生模型的损失(9)的恒等化简到了尽头了吗？并没有，因为它的内积左边还有$\boldsymbol{\epsilon}_{\boldsymbol{\psi}^*}(\boldsymbol{x}_t^{(g)}, t)$，还可以用同样的方式进行化简。具体来说，我们有

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{z},\boldsymbol{\varepsilon}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})}\left[\left\|\boldsymbol{\epsilon}_{\boldsymbol{\varphi}^*}(\boldsymbol{x}_t^{(g)},t)-\boldsymbol{\epsilon}_{\boldsymbol{\psi}^*}(\boldsymbol{x}_t^{(g)},t)\right\|^2\right]\\
=&\mathbb{E}_{\boldsymbol{z},\boldsymbol{\varepsilon}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})}\left[\left\|\boldsymbol{\epsilon}_{\boldsymbol{\varphi}^*}(\boldsymbol{x}_t^{(g)},t)\right\|^2-2\left\langle\boldsymbol{\epsilon}_{\boldsymbol{\varphi}^*}(\boldsymbol{x}_t^{(g)},t),\boldsymbol{\epsilon}_{\boldsymbol{\psi}^*}(\boldsymbol{x}_t^{(g)},t)\right\rangle+\left\|\boldsymbol{\epsilon}_{\boldsymbol{\psi}^*}(\boldsymbol{x}_t^{(g)},t)\right\|^2\right]\\
=&\mathbb{E}_{\boldsymbol{x}_t^{(g)}\sim p_{\boldsymbol{\theta}}(\boldsymbol{x}_t^{(g)})}\left[\left\|\boldsymbol{\epsilon}_{\boldsymbol{\varphi}^*}(\boldsymbol{x}_t^{(g)},t)\right\|^2-2\left\langle\boldsymbol{\epsilon}_{\boldsymbol{\varphi}^*}(\boldsymbol{x}_t^{(g)},t),-\bar{\beta}_t\nabla_{\boldsymbol{x}_t^{(g)}}\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_t^{(g)})\right\rangle+\left\langle\boldsymbol{\epsilon}_{\boldsymbol{\psi}^*}(\boldsymbol{x}_t^{(g)},t),-\bar{\beta}_t\nabla_{\boldsymbol{x}_t^{(g)}}\log p_{\boldsymbol{\theta}}(\boldsymbol{x}\right.\right.
\end{aligned}$$

这里的每一个$-\bar{\beta}_t\nabla_{\boldsymbol{x}_t^{(g)}}\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_t^{(g)})$都可以用相同的恒等变换(7)最终转化为单个$\boldsymbol{\varepsilon}$（但要注意$\|\boldsymbol{\epsilon}_{\boldsymbol{\psi}^*}(\boldsymbol{x}_t^{(g)},t)\|^2=\langle\boldsymbol{\epsilon}_{\boldsymbol{\psi}^*}(\boldsymbol{x}_t^{(g)},t),\boldsymbol{\epsilon}_{\boldsymbol{\psi}^*}(\boldsymbol{x}_t^{(g)},t)\rangle$只能转换一个，不能都转），而式(9)相当于只转了一部分，如果全部转会更好吗？因为没有实验结果，所以暂时不得而知。但有一个特别有意思的形式，就是只转换上面的中间部分的话，该损失函数可以写成

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{z},\boldsymbol{\varepsilon}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})}\left[\left\|\boldsymbol{\epsilon}_{\boldsymbol{\varphi}^*}(\boldsymbol{x}_t^{(g)},t)-\boldsymbol{\epsilon}_{\boldsymbol{\psi}^*}(\boldsymbol{x}_t^{(g)},t)\right\|^2\right]\\
=&\mathbb{E}_{\boldsymbol{z},\boldsymbol{\varepsilon}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})}\left[\left\|\boldsymbol{\epsilon}_{\boldsymbol{\varphi}^*}(\boldsymbol{x}_t^{(g)},t)\right\|^2-2\left\langle\boldsymbol{\epsilon}_{\boldsymbol{\varphi}^*}(\boldsymbol{x}_t^{(g)},t),\boldsymbol{\varepsilon}\right\rangle+\left\|\boldsymbol{\epsilon}_{\boldsymbol{\psi}^*}(\boldsymbol{x}_t^{(g)},t)\right\|^2\right] \qquad (11)\\
=&\mathbb{E}_{\boldsymbol{z},\boldsymbol{\varepsilon}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})}\left[\left\|\boldsymbol{\epsilon}_{\boldsymbol{\varphi}^*}(\boldsymbol{x}_t^{(g)},t)-\boldsymbol{\varepsilon}\right\|^2+\left\|\boldsymbol{\epsilon}_{\boldsymbol{\psi}^*}(\boldsymbol{x}_t^{(g)},t)\right\|^2\right]+\text{常数}
\end{aligned}$$

这是学生模型，也就是生成器的损失，然后我们再对比学生数据去噪模型的损失(3)：

$$\boldsymbol{\psi}^*=\mathop{\text{argmin}}_{\boldsymbol{\psi}}\mathbb{E}_{\boldsymbol{z},\boldsymbol{\varepsilon}\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I})}\left[\left\|\boldsymbol{\epsilon}_{\boldsymbol{\psi}}(\boldsymbol{x}_t^{(g)},t)-\boldsymbol{\varepsilon}\right\|^2\right] \qquad (12)$$

这两个式子联合起来看，我们可以发现学生模型实则在向教师模型看齐，并且试图远离学生数据所训练的去噪模型，形式上很像LSGAN，$\boldsymbol{\epsilon}_{\boldsymbol{\psi}}(\boldsymbol{x}_t^{(g)},t)$类似GAN的判别器，不同的地方是，GAN的判别器一般是两项损失相加而生成器是单项损失，SiD则反过来了。这其实体现了两种不同的学习思路：

> 1、GAN：一开始造假者（生成器）和鉴别者（判别器）都是小白，鉴别者不断对比真品和赝品来提供自己的鉴宝水平，造假者则通过鉴别者的反馈不断提高自己的造假水平；

> 2、SiD：完全没有真品，但有一个绝对权威的鉴宝大师（教师模型），造假者（学生模型）不断制作赝品，同时培养自己的鉴别者（学生数据训练的去噪模型），然后通过自家鉴别者跟大师的交流来提高自己造假水平。

可能有读者会问：为什么SiD中的造假者不直接向大师请教，而是要通过培养自己的鉴别者来间接获得反馈呢？这是因为直接跟大师交流的话，可能会出现的问题就是长期都只交流同一个作品的技术，最终只制造出了一种能够以假乱真的赝品（模式坍缩），而通过培养自己的鉴别者一定程度上就可以避免这个问题，因为造假者的学习策略是“多得到大师的好评，同时尽量减少自家人的好评”，如果造假者还是只制造一种赝品，那么大师和自家的好评都会越来越多，这不符合造假者的学习策略，从而迫使造假者不断开发新的产品而不是固步自封。

此外，读者可以发现，SiD整个训练并没有利用到扩散模型的递归采样的任何信息，换句话说它纯粹是利用了去噪这一训练方式所训练出来的去噪模型，那么一个自然的问题是：如果单纯为了训练一个单步的生成模型，而不是作为已有扩散模型的蒸馏，那么我们训练一个只具有单一噪声强度的去噪模型会不会更好？比如像DDE一样，固定$\bar{\alpha}_t = 1$、$\bar{\beta}_t = \beta =$某个常数取训练一个去噪模型，然后用它来重复SiD的训练过程，这样会不会能够简化训练难度、提高训练效率？这也是一个值得进一步确认的问题。

文章小结

在这篇文章中，我们介绍了一种新的将扩散模型蒸馏为单步生成模型的方案，其思想可以追溯到前两年的利用去噪自编码器训练生成模型的工作，它不需要获得教师模型的真实训练集，也不需要迭代教师模型来生成样本对，而引入了类似GAN的交替训练，同时提出了关键的恒等变换来稳定训练过程，整个方法有颇多值得学习之处。

如果您需要引用本文，请参考：

苏剑林. (May. 01, 2024). 《生成扩散模型漫谈（二十五）：基于恒等式的蒸馏（上）》[Blog post]. Retrieved from https://spaces.ac.cn/archives/10085

@online{kexuefm-10085,
title={生成扩散模型漫谈（二十五）：基于恒等式的蒸馏（上）},
author={苏剑林},
year={2024},
month={May},
url={\url{https://spaces.ac.cn/archives/10085}},
}