扩散模型过程整理_Nekasu

> ⚠ 提示
>
> 点击右上角「书本」图标, 进入阅读模式, 以
> 获得更好的阅读体验!

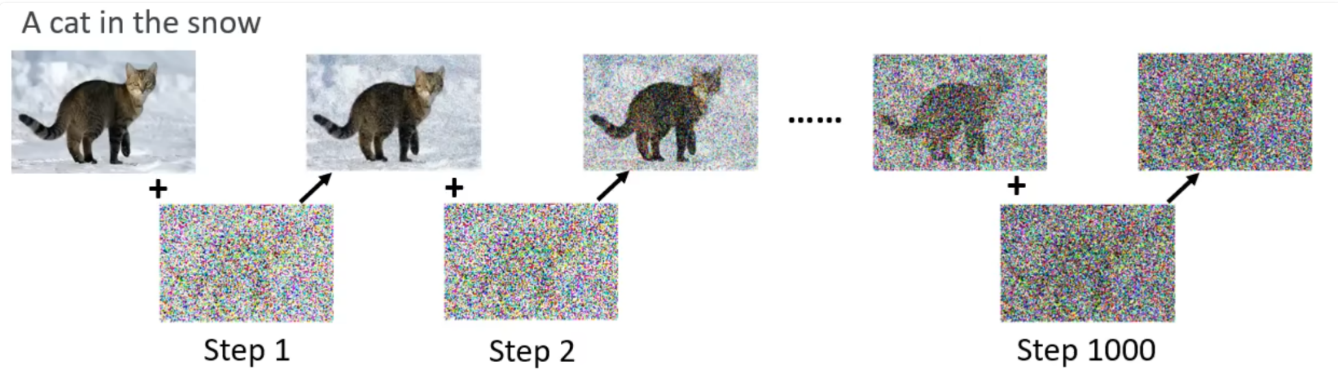
> *作者：Nekasu*
>
> *zxt1428147954@163.com*
>
> *1428147954@qq.com*

扩散模型的前提

1. 扩散模型遵循这样一个前提：所有图像都满足某种特定的概率分布
    1. 高斯噪声服从高斯分布, 且高斯噪声是从高斯分布中采样得到的图像
    2. 复杂图像服从复杂分布, 且复杂图像是从某个复杂分布中采样得到的图像
2. 如何将复杂分布变成高斯分布, 就是扩散过程需要完成的任务.
    1. 实际上, 这一步很简单
    2. 在实际中, 我们往往是有一张特定的图像, 只需要往里面加高斯噪声就可以达成这个目标
    3. 这是因为这个特定的图像可以看作是一个定值, 所以可以看作是高斯噪声的均值
3. 如何将高斯分布变换成其他复杂分布, 就是逆扩散过程需要完成的任务
    1. 经典的扩散模型就是从高斯高斯分布中, 一点一点减去复杂的高斯分布, 以达成这个目标的

前向过程与扩散

前向过程的描述



前向过程的参数化表示

参数设定

1. 原始图像：x_0
2. 第 t 次的从标准高斯分布中采样噪声图像 z_t
3. 第t次将噪声z_t加入x_0后的图像x_t
4. 第t次加噪声时, 噪声图像z_t与图像x_{t-1}的比例$1-\beta_t$与β_t

加噪过程

加噪过程公式

1. 从数据集中获取一张原始的真实图像x_0
2. 从标准高斯分布$\mathcal{N}(0,1)$中采样一张噪声图z_1
3. 将噪声图z_1与原始图像x_0按$\sqrt{1-\alpha_1}$与$\sqrt{\alpha_1}$的比例混合, 可以得到第一步的加噪声结果, 如下所示

$$x_1 = \sqrt{1-\alpha_1} z_1 + \sqrt{\alpha_1} x_0$$

4. 将噪声图z_2与上一步得到的结果z_1按$\sqrt{1-\alpha_2}$与$\sqrt{\alpha_2}$的比例混合, 如下所示

$$x_2 = \sqrt{1-\alpha_2} z_2 + \sqrt{\alpha_2} x_1 \tag{2}$$

5. 则第t张加噪图像x_t满足以下公式:

$$x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1-\alpha_t} z_t \tag{3}$$

加噪过程公式简化

整体简化

我们对第t张加噪图像x_t满足的公式$x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1-\alpha_t} z_t$进行如下变换:

$$\begin{aligned}
x_t &= \sqrt{\alpha_t} x_{t-1} + \sqrt{1-\alpha_t} z_t \\
\\
&= \sqrt{\alpha_t} \left(\sqrt{\alpha_{t-1}} x_{t-2} + \sqrt{1-\alpha_{t-1}} z_{t-1} \right) + \sqrt{1-\alpha_t} z_t \\
&= \sqrt{\alpha_t \alpha_{t-1}} x_{t-2} + \sqrt{\alpha_t (1-\alpha_{t-1})} z_{t-1} + \sqrt{1-\alpha_t} z_t \\
\\
&= \sqrt{\alpha_t \alpha_{t-1}} \left(\sqrt{\alpha_{t-2}} x_{t-3} + \sqrt{1-\alpha_{t-2}} z_{t-2} \right) + \sqrt{\alpha_t (1-\alpha_{t-1})} z_{t-1} + \sqrt{1-\alpha_t} z_t \\
&= \sqrt{\alpha_t \alpha_{t-1} \alpha_{t-2}} x_{t-3} + \sqrt{\alpha_t \alpha_{t-1} (1-\alpha_{t-2})} z_{t-2} + \sqrt{\alpha_t (1-\alpha_{t-1})} z_{t-1} + \sqrt{1-\alpha_t} z_t \\
\\
&= \cdots \\
&= \sqrt{\alpha_t \alpha_{t-1} \alpha_{t-2} \cdots \alpha_1} x_0 + \sqrt{\alpha_t \alpha_{t-1} \cdots \alpha_2 (1-\alpha_1)} z_1 \\
&\qquad + \sqrt{\alpha_t \alpha_{t-1} \cdots \alpha_3 (1-\alpha_2)} z_2 \\
&\qquad + \sqrt{\alpha_t \alpha_{t-1} \cdots \alpha_4 (1-\alpha_3)} z_3 \\
&\qquad + \sqrt{\alpha_t \alpha_{t-1} \cdots \alpha_5 (1-\alpha_4)} z_4 \\
&\qquad + \sqrt{\alpha_t \alpha_{t-1} \cdots \alpha_6 (1-\alpha_5)} z_5 \\
&\qquad + \sqrt{\alpha_t \alpha_{t-1} \cdots \alpha_7 (1-\alpha_6)} z_6 \\
&\qquad \vdots \\
&\qquad + \sqrt{\alpha_t (1-\alpha_{t-1})} z_{t-1} \\
&\qquad + \sqrt{1-\alpha_t} z_t \\
\\
&= \sqrt{\alpha_t \alpha_{t-1} \alpha_{t-2} \cdots \alpha_1} x_0 + \sum_{i=1}^{t} \sqrt{\alpha_t \alpha_{t-1} \cdots \alpha_{i+1} (1-\alpha_i)} z_i
\end{aligned}$$

通过上述公式, 我们便可以计算经过t次加噪后获得的图像x_t. 整个公式可以看作是两个部分, 其一为包含原图的「原图项」$\sqrt{\alpha_t \alpha_{t-1} \alpha_{t-2} \cdots \alpha_1} x_0$, 其二为包含$t$个噪声的「叠加噪声项」$\sum_{i=1}^{t} \sqrt{\alpha_t \alpha_{t-1} \cdots \alpha_{i+1} (1-\alpha_i)} z_i$

原图项的简化

实际上, 「原图项」是一个计算较为简单的项, 而后面的「叠加噪声项」是一个计算复杂的项. 现在我们探索如何简化「叠加噪声项」.

为了化简方便, 我们定义「原图项」$\sqrt{\alpha_t \alpha_{t-1} \alpha_{t-2} \cdots \alpha_1} x_0$的系数$\sqrt{\alpha_t \alpha_{t-1} \alpha_{t-2} \cdots \alpha_1}$为$\overline{\alpha_t}$, 有

$$\overline{\alpha_t}=\prod_{i=1}^{t}\alpha_i=\sqrt{\alpha_t\alpha_{t-1}\alpha_{t-2}\cdots\alpha_1}$$

如$\overline{\alpha_2}=\alpha_2\cdot\alpha_1$,$\overline{\alpha_4}=\alpha_4\cdot\alpha_3\cdot\alpha_2\cdot\alpha_1$

在此定义下, 将$\overline{\alpha_t}=\prod\limits_{i=1}^{t}\alpha_i$代入$x_t$中的「原图项」, 有

$$\begin{aligned}x_t&=\sqrt{\alpha_t\alpha_{t-1}\alpha_{t-2}\cdots\alpha_1}x_0+\sum_{i=1}^{t}\sqrt{\alpha_t\alpha_{t-1}\cdots\alpha_{i+1}(1-\alpha_i)}z_i\\&=\sqrt{\overline{\alpha}_t}x_0+\sum_{i=1}^{t}\sqrt{\alpha_t\alpha_{t-1}\cdots\alpha_{i+1}(1-\alpha_i)}z_i\end{aligned}\tag{6}$$

叠加噪声项的简化

因为我们希望整个「叠加噪声项」是一个简单的噪声, 所以我们希望最终x_t满足的表达式也具有$x_t=\sqrt{\overline{\alpha}_t}a+\sqrt{1-\overline{\alpha}_t}b$的形式, 即系数的平方和为1.

注意到「原图项」的系数现在为$\sqrt{\overline{\alpha}_t}$, 我们期望后面的「叠加噪声项」的系数为$\sqrt{1-\overline{\alpha}_t}$, 所以对「叠加噪声项」提取系数$\sqrt{1-\overline{\alpha}_t}$, 从而有以下变化

$$\begin{aligned}x_t&=\sqrt{\overline{\alpha}_t}x_0+\sum_{i=1}^{t}\sqrt{\alpha_t\alpha_{t-1}\cdots\alpha_{i+1}(1-\alpha_i)}z_i\\&=\sqrt{\overline{\alpha}_t}x_0+\quad\sqrt{1-\overline{\alpha}_t}\sum_{i=1}^{t}\frac{\sqrt{\alpha_t\alpha_{t-1}\cdots\alpha_{i+1}(1-\alpha_i)}}{\sqrt{1-\overline{\alpha}_t}}z_i\end{aligned}$$

现在考察「叠加噪声项」中的$\sum_{i=1}^{t}\frac{\sqrt{\alpha_t\alpha_{t-1}\cdots\alpha_{i+1}(1-\alpha_i)}}{\sqrt{1-\overline{\alpha}_t}}z_i$部分 (即上面公式中最后一个分式)分以期望他是一个简单的分布.

这个分式部分中的单个噪声项z_i均是从标准高斯分布中采样得到的, 即$z_i\sim\mathcal{N}(0,1)$(即均值为0，方差为1), 且上一次采样的噪声不会影响下一次噪声的采样, 所以z_i的获取是相互独立的.

同时, 我们知道, 如果$X\sim\mathcal{N}(\mu_x,\sigma_x^2)$,$Y\sim\mathcal{N}(\mu_y,\sigma_y^2)$, 且$X$与$Y$相互独立, 则有$aX+bY\sim\mathcal{N}(a\mu_x+b\mu_y,a^2\sigma_x^2+b^2\sigma_y^2)$

在这样的情况下, 则有「叠加噪声项」中的$\sum_{i=1}^{t}\frac{\sqrt{\alpha_t\alpha_{t-1}\cdots\alpha_{i+1}(1-\alpha_i)}}{\sqrt{1-\overline{\alpha}_t}}z_i$部分服从以下高斯分布：

$$\sum_{i=1}^{t}\frac{\sqrt{\alpha_t\alpha_{t-1}\cdots\alpha_{i+1}(1-\alpha_i)}}{\sqrt{1-\overline{\alpha}_t}}z_i\sim\mathcal{N}\left[\sum_{i=1}^{t}\left(\frac{\sqrt{\alpha_t\alpha_{t-1}\cdots\alpha_{i+1}(1-\alpha_i)}}{\sqrt{1-\overline{\alpha}_t}}\cdot 0\right),\sum_{i=1}^{t}\left(\frac{\alpha_t\alpha_{t-1}\cdots\alpha_{i+1}(1-\alpha_i)}{1-\overline{\alpha}_t}\cdot 1\right)\right]$$

可以发现, 这个高斯分布的均值部分为0, 因为均值部分中累加的每一项都与0相乘了. 从而上述高斯分布可以化简为:

$$\sum_{i=1}^{t}\frac{\sqrt{\alpha_t\alpha_{t-1}\cdots\alpha_{i+1}(1-\alpha_i)}}{\sqrt{1-\overline{\alpha}_t}}z_i\sim\mathcal{N}\left(0,\sum_{i=1}^{t}\frac{\alpha_t\alpha_{t-1}\cdots\alpha_{i+1}(1-\alpha_i)}{1-\overline{\alpha}_t}\right)$$

为了搞清楚这个「叠加噪声项」到底满足什么样的高斯分布, 我们继续考察这个高斯分布的方差部分, 记为$\sigma_s^2=\sum_{i=1}^{t}\frac{\alpha_t\alpha_{t-1}\cdots\alpha_{i+1}(1-\alpha_i)}{1-\overline{\alpha}_t}$, 下标$s$为 sum 的缩写, 表示累加

直接将σ_S^2拆开, 有

$$
\begin{aligned}
\sigma_s^2 &= \sum_{i=1}^{t} \frac{\alpha_t\alpha_{t-1}\cdots\alpha_{i+1}(1-\alpha_i)}{1-\overline{\alpha_t}} \\
&= \frac{1}{1-\overline{\alpha_t}}\left[\sum_{i=1}^{t}\alpha_t\alpha_{t-1}\cdots\alpha_{i+1}(1-\alpha_i)\right] \\
&\xlongequal{\text{去除累加符号}} \frac{1}{1-\overline{\alpha_t}}[(1-\alpha_t)+\alpha_t(1-\alpha_{t-1})+\alpha_t\alpha_{t-1}(1-\alpha_{t-2})+\cdots+\alpha_t\alpha_{t-1}\cdots\alpha_2(1-\alpha_1)] \\
&\xlongequal{\text{拆除小括号}} \frac{1}{1-\overline{\alpha_t}}[1-\alpha_t+\alpha_t-\alpha_t\alpha_{t-1}+\alpha_t\alpha_{t-1}-\alpha_t\alpha_{t-1}\alpha_{t-2}+\cdots+\alpha_t\alpha_{t-1}\cdots\alpha_2-\alpha_t\alpha_{t-1}\cdots\alpha_2\alpha_1] \\
&\xlongequal{\text{发现中括号中除了第一项和最后一项都可以消去}} \frac{1}{1-\overline{\alpha_t}}[1-\alpha_t\alpha_{t-1}\cdots\alpha_2\alpha_1] \\
&= \frac{1}{1-\overline{\alpha_t}}(1-\overline{a_t}) \\
&\xlongequal{\text{分子分母相同}} 1
\end{aligned}
$$

也即$\sigma_s^2=1$, 从而有$\sum_{i=1}^{t}\frac{\sqrt{\alpha_t\alpha_{t-1}\cdots\alpha_{i+1}(1-\alpha_i)}}{\sqrt{1-\overline{\alpha_t}}}z_i \sim \mathcal{N}(0,1)$, 我们记$\sum_{i=1}^{t}\frac{\sqrt{\alpha_t\alpha_{t-1}\cdots\alpha_{i+1}(1-\alpha_i)}}{\sqrt{1-\overline{\alpha_t}}}z_i = \tilde{z}$, 从而有$\tilde{z}\sim\mathcal{N}(0,1)$

至此, 我们可以发现, 「叠加噪声项」也是一个服从标准高斯分布的噪声, 从而我们可以得到第t步图像x_t与原图x_0之间的关系:

$$
\begin{aligned}
x_t &= \sqrt{\overline{\alpha_t}}x_0 + \sum_{i=1}^{t}\sqrt{\alpha_t\alpha_{t-1}\cdots\alpha_{i+1}(1-\alpha_i)}z_i \\
&= \sqrt{\overline{\alpha_t}}x_0 + \sqrt{1-\overline{a_t}}\tilde{z}, \quad \text{其中}\tilde{z}\sim\mathcal{N}(0,1)
\end{aligned}
$$

加噪过程总结

在公式$x_t=\sqrt{\overline{\alpha_t}}x_0+\sqrt{1-\overline{a_t}}\tilde{z}, \quad \text{其中}\tilde{z}\sim\mathcal{N}(0,1)$的指导下, 可以立刻得到某个特定步骤$t$的加噪图像$x_t$, 且加噪图像仅与原始图像$x_0$与当前步骤数$t$有关.

经过这样的简化, 就可以简单的获取加噪后的图像, 也即训练数据了.

概率采样视角看加噪过程

在上面的扩散模型过程整理_Nekasu > 加噪过程中, 我们推导出了直接从原图x_0获取第t次后的加噪图像x_t的公式如下:

$$
x_t=\sqrt{\overline{\alpha_t}}x_0+\sqrt{1-\overline{a_t}}\tilde{z}, \quad \text{其中}\tilde{z}\sim\mathcal{N}(0,1)
$$

实际上, 我们也可以将x_t看作是某种概率的采样结果, 推导如下:

已知$\tilde{z}\sim\mathcal{N}(0,1)$, 则有$\sqrt{1-\overline{a_t}}\tilde{z}\sim\mathcal{N}(0,1-\overline{\alpha_t})$

从而有$\sqrt{\overline{\alpha_t}}x_0+\sqrt{1-\overline{a_t}}\tilde{z}\sim\mathcal{N}(\sqrt{\overline{\alpha_t}}x_0, 1-\overline{\alpha_t})$,

从而有$q(x_t|x_0)\sim\mathcal{N}(\sqrt{\overline{\alpha_t}}x_0, 1-\overline{\alpha_t})$

这个结果表明，第t步的加噪图像x_t 可以看作是从一个正态分布中采样的结果，其均值和方差分别由初始图像x_0和累积噪声参数$\overline{\alpha_t}$决定

同理, 由于有$x_t=\sqrt{\alpha_t}x_{t-1}+\sqrt{1-\alpha_t}z_t$, 所以有

$$
q(x_t|x_{t-1},x_0)\sim\mathcal{N}(\sqrt{\alpha_t}x_{t-1}, 1-\alpha_t)
$$

逆向过程与去噪

逆向过程的描述

从后一个推前一个

逆向过程的符号表示

很显然, 我们想要通过x_t来预测x_{t-1}.

如果我们能够逆转上述扩散过程, 并从$p(x_{t-1}|x_t)$采样，就可以从高斯噪声$x_t \sim N(0,1)$还原出原图服从的分布$x_0 \sim p(x)$。

如何获得$p_\theta(x_{t-1}|x_t)$这个概率密度 (式子中的θ表示神经网络的参数)就是一个需要探讨的问题. 直接计算比较困难, 所以我们可以考虑对公式进行变形. 对公式使用贝叶斯公式, 有如下结果

$$\begin{aligned} p_\theta(x_{t-1}|x_t) &\xlongequal{\text{贝叶斯公式}} \frac{p_\theta(x_{t-1},x_t)}{p(x_t)} \\ &\xlongequal{\text{分母用全概率公式展开}} \frac{p(x_t|x_{t-1})\cdot p_\theta(x_{t-1})}{p(x_t)} \\ &= p(x_t|x_{t-1})\cdot\frac{p_\theta(x_{t-1})}{p(x_t)} \end{aligned}$$

通过这样的变换, 我们将一个无法计算的式子$p_\theta(x_{t-1}|x_t)$, 改写成了一个可计算的部分$p(x_t|x_{t-1})$和一个不可计算的分式$\frac{p_\theta(x_{t-1})}{p(x_t)}$的乘积

自己观察这个不可计算的分式, 可以发现, 这个分式的分子与分母都是不可计算的. 因为如果我们能直接得到$p(x_t)$或$p(x_{x-1})$, 那我们就能直接得出$p(x_0)$. 但是我们计算$p(x_{t-1}|x_t)$的目的就是为了计算$p(x_0)$, 如果可以直接得出$p(x_0)$, 那么我们就没有计算$p(x_{t-1}|x_t)$, 所以我们不可能直接得到$p(x_t)$.

可以想到, 虽然直接计算$p(x_t)$是不可行的, 但是计算$p(x_t|x_0)$是十分简单的, 在扩散模型过程整理_Nekasu > 概率采样视角看加噪过程 中我们介绍过这个计算:

$$p(x_t|x_0) \sim \mathcal{N}(\sqrt{\overline{\alpha}_t}x_0, 1-\overline{\alpha_t})$$

所以我们可以考虑将$p(x_t)$的计算转换成$p(x_t|x_0)$用于计算, 由此可以计算$p_\theta(x_{t-1}|x_t,x_0)$, 有

$$\begin{aligned} p_\theta(x_{t-1}|x_t,x_0) &\xlongequal{\text{贝叶斯公式}} \frac{p_\theta(x_{t-1},x_t,x_0)}{p(x_t,x_0)} \\ &\xlongequal{\text{分子用全概率公式展开}} \frac{p(x_t|x_{t-1},x_0)\cdot p_\theta(x_{t-1},x_0)}{p(x_t,x_0)} \\ &= p(x_t|x_{t-1})\cdot\frac{p(x_{t-1},x_0)}{p(x_t,x_0)} \\ &= p(x_t|x_{t-1})\cdot\frac{p(x_{t-1},x_0)}{p(x_t,x_0)} \\ &=p(x_t|x_{t-1})\cdot\frac{p(x_{t-1}|x_0)\cdot p(x_0)}{p(x_t|x_0)\cdot p(x_0)} \\ &=p(x_t|x_{t-1})\cdot\frac{p(x_{t-1}|x_0)}{p(x_t|x_0)} \end{aligned}$$

对上述推导取等式左侧与右侧第一项, 有 $p_\theta(x_{t-1}|x_t,x_0) = p(x_t|x_{t-1})\cdot\frac{p_\theta(x_{t-1}|x_0)}{p(x_t|x_0)}$

可能有人发现, 在第三个等式中, 我们直接将 $p(x_t|x_{t-1},x_0)$ 替换为了 $p(x_t|x_{x-1})$, 这可以用马尔可夫性（Markov property）来解释：马尔可夫性假设指出，一个状态只依赖于前一个状态，而与更早的状态条件独立。应用于 DDPM 模型，这意味着：

$$p(x_t \mid x_{t-1},x_0) = p(x_t \mid x_{t-1})p(x_t|x_{t-1},x_0) = p(x_t|x_{t-1})p(x_t \mid x_{t-1},x_0) = p(x_t \mid x_{t-1})$$

这三项就都是很好计算的项了, 从扩散模型过程整理_Nekasu > 概率采样视角看加噪过程 中, 虽然此处我们使用的是$p(x_t|x_{t-1})$而非$q(x_t|x_{t-1})$, 但他们之间表示的内容是一致的, 均表示前向过程中的加噪.

我们有推导结果

$$
\begin{aligned}
p(x_t|x_{t-1},x_0)=q(x_t|x_{t-1},x_0)\sim\mathcal{N}(\sqrt{\alpha_t}x_{t-1},1-\alpha_t)\\
&=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\cdot(\frac{x_t-\mu}{\sigma})^2}\\
&=\frac{1}{\sqrt{2\pi}\cdot\sqrt{1-\alpha_t}}\exp\left[-\frac{1}{2}\cdot\left(\frac{x_t-\sqrt{\alpha_t}x_{t-1}}{\sqrt{1-\alpha_t}}\right)^2\right]
\end{aligned}
$$

x_t为随机变量

$$
\begin{aligned}
p_\theta(x_{t-1}|x_0)=q(x_{t-1}|x_0)\sim\mathcal{N}(\sqrt{\overline{\alpha_t}}x_0,1-\overline{\alpha_{t-1}})\\
&=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\cdot(\frac{x_{t-1}-\mu}{\sigma})^2}\\
&=\frac{1}{\sqrt{2\pi}\cdot\sqrt{1-\overline{\alpha_{t-1}}}}\exp\left[-\frac{1}{2}\cdot\left(\frac{x_{t-1}-\sqrt{\overline{\alpha_{t-1}}}x_0}{\sqrt{1-\overline{\alpha_{t-1}}}}\right)^2\right]
\end{aligned}
$$

x_{t-1}为随机变量

$$
\begin{aligned}
p(x_t|x_0)=q(x_t|x_0)\sim\mathcal{N}(\sqrt{\overline{\alpha_t}}x_0,1-\overline{\alpha_t})\\
&=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\cdot(\frac{x_t-\mu}{\sigma})^2}\\
&=\frac{1}{\sqrt{2\pi}\cdot\sqrt{1-\overline{\alpha_t}}}\exp\left[-\frac{1}{2}\cdot\left(\frac{x_t-\sqrt{\overline{\alpha_t}}x_0}{\sqrt{1-\overline{\alpha_t}}}\right)^2\right]
\end{aligned}
$$

x_t为随机变量

将这三个式子带入上述公式$p_\theta(x_{t-1}|x_t,x_0)=p(x_t|x_{t-1})\cdot\frac{p_\theta(x_{t-1}|x_0)}{p(x_t|x_0)}$, 有如下化简:

$$
\begin{aligned}
p_\theta(x_{t-1}|x_t,x_0)&=p(x_t|x_{t-1})\cdot\frac{p_\theta(x_{t-1}|x_0)}{p(x_t|x_0)}\\
&=\frac{1}{\sqrt{2\pi}\cdot\sqrt{1-\alpha_t}}\exp\left[-\frac{1}{2}\cdot\left(\frac{x_t-\sqrt{\alpha_t}x_{t-1}}{\sqrt{1-\alpha_t}}\right)^2\right]\cdot\frac{\frac{1}{\sqrt{2\pi}\cdot\sqrt{1-\overline{\alpha_{t-1}}}}\exp\left[-\frac{1}{2}\cdot\left(\frac{x_{t-1}-\sqrt{\overline{\alpha_{t-1}}}x_0}{\sqrt{1-\overline{\alpha_{t-1}}}}\right)^2\right]}{\frac{1}{\sqrt{2\pi}\cdot\sqrt{1-\overline{\alpha_t}}}\exp\left[-\frac{1}{2}\cdot\left(\frac{x_t-\sqrt{\overline{\alpha_t}}x_0}{\sqrt{1-\overline{\alpha_t}}}\right)^2\right]}\\
&\xlongequal[\text{忽略前面的系数}]{\text{指数相乘(除)等于幂相加(减)}}k\cdot\exp\left\{-\frac{1}{2}\cdot\left[\left(\frac{x_t-\sqrt{\alpha_t}x_{t-1}}{\sqrt{1-\alpha_t}}\right)^2+\left(\frac{x_{t-1}-\sqrt{\overline{\alpha_{t-1}}}x_0}{\sqrt{1-\overline{\alpha_{t-1}}}}\right)^2-\left(\frac{x_t-\sqrt{\overline{\alpha_t}}x_0}{\sqrt{1-\overline{\alpha_t}}}\right)^2\right]\right\}
\end{aligned}
$$

这个公式看起来很复杂, 但是我们可以这样考虑：整个公式为一个常数与一个「以e为底的指数」的乘积, 与高斯分布的形式很像, 而我们知道, 高斯分布中的指数部分是一个完全平方：
$\exp\left(-\frac{1}{2}\cdot\left[\frac{(x-\mu)}{\sigma}\right]^2\right)=\exp\left\{-\frac{1}{2}\left(\frac{1}{\sigma^2}x^2-\frac{2\mu}{\sigma^2}x+\frac{\mu^2}{\sigma^2}\right)\right\}$, 所以我们可以也将上面公式中的指数部分变成一个完全平方.

经过简化, 有:

$$
\begin{aligned}
p_\theta(x_{t-1}|x_t,x_0) &= k\cdot\exp\left\{-\frac{1}{2}\cdot\left[\left(\frac{x-\sqrt{\alpha_t}x_{t-1}}{\sqrt{1-\alpha_t}}\right)^2+\left(\frac{x-\sqrt{\overline{\alpha}_{t-1}}x_0}{\sqrt{1-\overline{\alpha}_{t-1}}}\right)^2-\left(\frac{x-\sqrt{\overline{\alpha}_t}x_0}{\sqrt{1-\overline{\alpha}_t}}\right)^2\right]\right\}\\
&\xlongequal{将平方拆开} k\cdot\exp\left\{-\frac{1}{2}\left(\frac{x_t^2-2\sqrt{\alpha_t}x_tx_{t-1}+\alpha_tx_{t-1}^2}{1-\alpha_t}+\frac{x_{t-1}^2-2\sqrt{\overline{\alpha}_{t-1}}x_0x_{t-1}+\overline{\alpha}_{t-1}x_0^2}{1-\overline{\alpha}_{t-1}}-\frac{\left(x_t-\sqrt{\overline{\alpha}_t}x_0\right)}{1-\overline{\alpha}_t}\right.\right.\\
&\xlongequal{合并x_{t-1}的同类项} k\cdot\exp\left\{\underbrace{-\frac{1}{2}\left[\left(\frac{\alpha_t}{1-\alpha_t}+\frac{1}{1-\overline{\alpha}_{t-1}}\right)x_{t-1}^2-2\left(\frac{\sqrt{\alpha_t}}{1-\alpha_t}x_t+\frac{\sqrt{\overline{\alpha}_{t-1}}}{1-\overline{\alpha}_{t-1}}x_0\right)x_{t-1}+C\left(x_t,x_0\right)\right.}_{这部分就是高斯分布中的-\frac{1}{2}\left(\frac{1}{\sigma^2}x^2-\frac{2\mu}{\sigma^2}x+\frac{\mu^2}{\sigma^2}\right)}\right.\\
&\xlongequal{完全平方公式} k\cdot\exp\left\{-\frac{1}{2}\cdot\left[\frac{x_{t-1}-\left(\frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})}{1-\overline{\alpha}_t}x_t+\frac{\sqrt{\overline{\alpha}_{t-1}}(1-\alpha_t)}{1-\overline{\alpha}_t}x_0\right)}{\sqrt{\frac{\alpha_t}{1-\alpha_t}+\frac{1}{1-\overline{\alpha}_{t-1}}}}\right]^2\right\}\\
&= \exp\left(-\frac{1}{2}\cdot\left[\frac{(x-\mu)}{\sigma}\right]^2\right)
\end{aligned}
$$

其中,$\mu=\frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})}{1-\overline{\alpha}_t}x_t+\frac{\sqrt{\overline{\alpha}_{t-1}}(1-\alpha_t)}{1-\overline{\alpha}_t}x_0,\sigma=\sqrt{\frac{\alpha_t}{1-\alpha_t}+\frac{1}{1-\overline{\alpha}_{t-1}}}$

通过以上化简, 我们轻易的得到的$p_\theta(x_{t-1}|x_t,x_0)$所服从的分布实际上也是一个高斯分布, 即

$$
p_\theta(x_{t-1}|x_t,x_0)\sim\mathcal{N}(\mu,\sigma^2)\sim\mathcal{N}\left(\frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})}{1-\overline{\alpha}_t}x_t+\frac{\sqrt{\overline{\alpha}_{t-1}}\,(1-\alpha_t)}{1-\overline{\alpha}_t}x_0,\frac{\alpha_t}{1-\alpha_t}+\frac{1}{1-\overline{\alpha}_{t-1}}\right)
$$

观察方差$\sigma^2=\frac{\alpha_t}{1-\alpha_t}+\frac{1}{1-\overline{\alpha}_{t-1}}$可以发现, 其中所有的值均为常数, 是一个可以直接计算的值

观察均值$\mu=\frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})}{1-\overline{\alpha}_t}x_t+\frac{\sqrt{\overline{\alpha}_{t-1}}(1-\alpha_t)}{1-\overline{\alpha}_t}x_0$, 可以发现所有的$\alpha$均是已知值,$x_t$也是已知值, 而整个均值的式子中, 唯一一个不知道的值为x_0. 如果能得知x_0就能很快的进行计算了.

实际上, 我们在扩散模型过程整理_Nekasu > 加噪过程总结 中有公式

$$
x_t=\sqrt{\overline{\alpha}_t}x_0+\sqrt{1-\overline{\alpha}_t}\tilde{z},\quad 其中\tilde{z}\sim\mathcal{N}(0,1)
$$

从这个式子中, 我们可以反推出x_0, 即$x_0=\frac{1}{\sqrt{\overline{\alpha}_t}}\left(x_{t-1}-\sqrt{1-\overline{\alpha}_t}\tilde{z}\right)$, 用这个值替换$\mu$中的$x_0$, 有如下结果:

$$
\begin{aligned}
\mu &= \frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})}{1-\overline{\alpha}_t}x_t+\frac{\sqrt{\overline{\alpha}_{t-1}}\,(1-\alpha_t)}{1-\overline{\alpha}_t}x_0\\
&= \frac{\sqrt{\alpha_t}(1-\overline{\alpha}_{t-1})}{1-\overline{\alpha}_t}x_t+\frac{\sqrt{\overline{\alpha}_{t-1}}\,(1-\alpha_t)}{1-\overline{\alpha}_t}\cdot\frac{1}{\sqrt{\overline{\alpha}_t}}\left(x_{t-1}-\sqrt{1-\overline{\alpha}_t}\tilde{z}\right)\\
&= \frac{1}{\sqrt{\alpha_t}}\left(x_t-\frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\tilde{z}\right)
\end{aligned}
$$

即

$$
\mu=\frac{1}{\sqrt{\alpha_t}}\left(x_t-\frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\tilde{z}\right)
$$

整个式子中仅有$\tilde{z}$是一个不可直接获得的值了. 为何会出现一个不能直接获得的$\tilde{z}$呢? 回顾上述过程发现, 这个$\tilde{z}$是在为了消去x_0时引入的一个值. 这个$\tilde{z}$表示的是在获得x_t的过程中, 向x_0中加入的噪声.

为了能够获得这个噪声, 原文使用了一个神经网络进行预测.

在使用神经网络获得$\tilde{z}$后,μ与σ^2都变的可以计算, 从而$p_\theta(x_{t-1}|x_t,x_0)\sim\mathcal{N}(\mu,\sigma^2)$就变成了可以直接获得的高斯分布. 从而$x_{t-1}$就可以轻松的从这个分布中采样获取了.

噪声预测器训练过程

在 DDPM 中, 使用 UNet 网络进行噪声图像的预测, 并采样极大似然函数作为损失函数进行预测.

首先我们介绍一下极大似然函数与极大似然估计

极大似然函数与估计

大家可能对极大似然估计有一些印象. 实际上, 这个名词在概率论与数理统计中讲过, 是属于数理统计的部分.

想要了解极大似然估计, 请看视频十分钟搞定极大似然估计.

利用极大似然函数进行参数迭代

我们知道, 极大似然估计在参数估计时, 使用的是这样的想法：小概率事件在现实中几乎不发生. 所以如果我们观测到某种事件发生了, 则说明这个事件所服从的分布中, 已经发生的这个事件一定是一个大概率事件. 能将所有已发生的事件的概率都最大化的参数, 也就是使似然函数最大的参数, 则是我们需要的参数.

从另一个角度说, 如果似然函数 L 越大, 则说明当前取得的参数越合理.

所以我们使用似然函数 L 作为似然函数, 将最大化似然函数 $\max L$ 作为训练的目标.

复习一下极大似然函数的构建过程

我们先用一个简单例子复习一下极大似然函数的构建过程. 如果你对极大似然估计很熟悉的话, 可以跳过这一部分.

如果下一部分中构建似然函数的过程难以理解, 可以与这一个简单的例子进行类比.

1. 现在有这样一个场景
    1. 我们有一个装了不知道多少个黑白小球的袋子 (是一个满足二项分布的模型)
        1. 取得黑球的概率为 θ, 取得白球的概率为 $1-\theta$
    2. 现在进行 5 次采样, 有结果如下：
        1. 采样 1：黑
        2. 采样 2：黑
        3. 采样 3：黑
        4. 采样 4：白
        5. 采样 5：白
    3. 问这个二项分布的参数 θ 是多少
2. 采样与二项分布
    1. 对于采样 1, 摸到了黑球, 在二项分布中有概率为 θ
    2. 对于采样 2, 摸到了黑球, 在二项分布中有概率为 θ
    3. 对于采样 3, 摸到了黑球, 在二项分布中有概率为 θ
    4. 对于采样 4, 摸到了白球, 在二项分布中有概率为 $1-\theta$
    5. 对于采样 5, 摸到了白球, 在二项分布中有概率为 $1-\theta$
    6. 对于整个五次采样结果为 (黑、黑、黑、白、白) 的概率为：$\theta^3\cdot(1-\theta)^2$
3. 极大似然估计的思想
    1. 既然五次采样出现了这样的结果, 所以我们认为发生这种 (黑、黑、黑、白、白) 情况的概率应该是最大的 (因为小概率事件不可能发生)
4. 极大似然估计
    1. 所以我们认为, θ 一定能使采样概率 $\theta^3\cdot(1-\theta)^2$ 取到最大值.
    2. 也即, 一个能使概率 $\theta^3\cdot(1-\theta)^2$ 达到最大值的 θ 是一个合理的 θ
    3. 所以应该求 $\theta^3\cdot(1-\theta)^2$ 取最大值时, θ 的取值.
    4. 因为 $\theta^3\cdot(1-\theta)^2$ 是一个高次幂的式子, 难以计算, 我们我们转而计算 $L(\theta)=\log\left[\theta^3\cdot(1-\theta)^2\right]$
    5. $L(\theta)=\log\left[\theta^3\cdot(1-\theta)^2\right]$ 即为似然函数

扩散模型的场景与极大似然估计

当我们复习了一个简单情形的似然估计的用法后, 可以快速的类比到当前的扩散模型任务中.

1. 现在我们有扩散模型的场景
    1. 我们有一个装了不知道多少个各异图像的图像集 (是一个满足 $p_\theta(x_{t-1}|x_t)$ 的模型)
        1. 取得图像 x_t 的概率为 $p_\theta(x_{t-1}|x_t)$, $t=1,2,3,\cdots,T-1$
        2. 取得图形 x_T 的概率为 $p(x_T)\sim\mathcal{N}(0,1)$
    2. 现在进行 T 次采样, 有结果如下
        1. 采样 0：x_0
        2. 采样 1：x_1
        3. 采样 2：x_3
        4. $\cdots$
        5. 采样 t：x_t
        6. $\cdots$
        7. 采样 T：x_T
    3. 问这个概率模型 $p_\theta(x_t|x_{t+1})$ 中的 θ 是多少
2. 采样与二项分布
    1. 对于采样 0, 获得了图像 x_0, 在分布中有概率 $p_\theta(x_0|x_1)$.
    2. 对于采样 1, 获得了图像 x_1, 在分布中有概率 $p_\theta(x_1|x_2)$.
    3. $\cdots$
    4. 对于采样 t, 获得了图像 x_t, 在分布中有概率 $p_\theta(x_t|x_{t+1})$.
    5. $\cdots$
    6. 对于采样 $T-1$, 获得了图像 x_{T-1}, 在分布中有概率 $p_\theta(x_{T-1}|x_T)$
    7. 对于采样 T, 获得了图像 x_T, 在分布中有概率 $p(x_T)\sim\mathcal{N}(0,1)$.
        1. 因为 x_T 是直接从标准高斯分布中获取的, 所以下标中没有 θ
    8. 对于整个 T 次采样结果为 $(x_0,x_1,x_2,\cdots,x_T)$ 的概率为 $p(x_T)\cdot p_\theta(x_{T-1}|x_T)\cdot p_\theta(x_{T-2}|x_{T-1})\cdots p_\theta(x_1|x_2)\cdot p(x_0|x_1)$
3. 极大似然估计的思想
    1. 既然 T 次采样出现了这样的结果, 所以我们认为发生这种 $(x_0,x_1,x_2,\cdots,x_T)$ 情况的概率应该是最大的 (因为小概率事件不可能发生)
4. 极大似然估计
    1. 所以我们认为, θ 一定能使采样概率 $p(x_T)\cdot p_\theta(x_{T-1}|x_T)\cdot p_\theta(x_{T-2}|x_{T-1})\cdots p_\theta(x_1|x_2)\cdot p_\theta(x_0|x_1)$ 取到最大值.
    2. 也即, 一个能使上述概率达到最大值的 θ 是一个合理的 θ
    3. 所以应该求 $p(x_T)\cdot p_\theta(x_{T-1}|x_T)\cdot p_\theta(x_{T-2}|x_{T-1})\cdots p_\theta(x_1|x_2)\cdot p_\theta(x_0|x_1)$ 取最大值时, θ 的取值.
    4. 因为 $p(x_T)\cdot p_\theta(x_{T-1}|x_T)\cdot p_\theta(x_{T-2}|x_{T-1})\cdots p_\theta(x_1|x_2)\cdot p_\theta(x_0|x_1)$ 是一个高次幂的式子, 难以计算, 我们我们转而计算 $L(\theta)=\log\left[p(x_T)\cdot p_\theta(x_{T-1}|x_T)\cdot p_\theta(x_{T-2}|x_{T-1})\cdots p_\theta(x_1|x_2)\cdot p_\theta(x_0|x_1)\right]$
    5. $L(\theta)=\log\left[p(x_T)\cdot p_\theta(x_{T-1}|x_T)\cdot p_\theta(x_{T-2}|x_{T-1})\cdots p_\theta(x_1|x_2)\cdot p_\theta(x_0|x_1)\right]$ 即为似然函数

实际上, 我们不可能简单的仅仅采样一个 x_0, 数据集中的所有图像都应该是一个可能的 x_0

所以我们要对式子 $p(x_T)\cdot p_\theta(x_{T-1}|x_T)\cdot p_\theta(x_{T-2}|x_{T-1})\cdots p_\theta(x_1|x_2)\cdot p(x_0|x_1)$ 进行积分, 有如下结果

$$P_\theta(x_0)=\int\limits_{x_1:x_T} p(x_T)\cdot p_\theta(x_{T-1}|x_T)\cdot p_\theta(x_{T-2}|x_{T-1})\cdots p_\theta(x_1|x_2)\cdot p_\theta(x_0|x_1)dx_1:x_T$$

所以有对数似然函数

$$L(\theta)=\log P_\theta(x_0)=\log\left(\int\limits_{x_1:x_T} p(x_T)\cdot p_\theta(x_{T-1}|x_T)\cdot p_\theta(x_{T-2}|x_{T-1})\cdots p_\theta(x_1|x_2)\cdot p_\theta(x_0|x_1)dx_1:x_T\right)$$

但实际上, 这个似然函数的计算比较困难, 所以在训练时不会直接进行计算.

那么该如何计算这个似然函数呢？请看下面的部分.

最大化似然函数的下界 (变分推断)

由于 DDPM 的过程有些复杂, 一共经过了 t 次的去噪, 比较复杂. 所以我们构建一个只有一个加噪过程的 DDPM 模型用来推导 (实际上这个「只有一个加噪过程的 DDPM」就是变分自编码器 VAE)

VAE 中的最大化下界

VAE 的对数似然函数

由于只有一个加噪过程, 所以有似然函数如下:

$$\log P_\theta(x) = \log \int_{x_1} p(x_1) p_\theta(x_0|x_1) dx_1$$

VAE 引入前向过程

从而可以有以下推导:

$$\begin{aligned}
\log P_\theta(x) &= \log \int_{x_1} p(x_1) p_\theta(x_0|x_1) dx_1 \\
&= \log \int_{x_1} p_\theta(x_0, x_1) dx_1 \\
&= \log \int_{x_1} \frac{q(x_1|x_0)}{q(x_1|x_0)} p_\theta(x_0, x_1) dx_1 \\
&= \log \int_{x_1} q(x_1|x_0) \frac{p_\theta(x_0, x_1)}{q(x_1|x_0)} dx_1 \\
&= \log \mathbb{E}_{q(x_1|x_0)} \left[\frac{p_\theta(x_0, x_1)}{q(x_1|x_0)} \right] \\
&\geq \mathbb{E}_{q(x_1|x_0)} \left[\log \frac{p_\theta(x_0, x_1)}{q(x_1|x_0)} \right] \qquad [\text{琴生不等式, 如果函数}\varphi\text{为凹函数, 则有}\varphi\left(\mathbb{E}(x)\right) \geq \mathbb{E}\left(\varphi(x)\right)]
\end{aligned}$$

VAE 中分离θ与常数

我们将 $p_\theta(x_0, x_1)$ 替换为 $p_\theta(x_0|x_1)p(x_1)$, 则有

$$\begin{aligned}
\mathbb{E}_{q(x_1|x_0)} \left[\log \frac{p_\theta(x_0, x_1)}{q(x_1|x_0)} \right] &= \mathbb{E}_{q(x_1|x_0)} \left[\log \frac{p_\theta(x_0|x_1)p(x_1)}{q(x_1|x_0)} \right] \\
&= \mathbb{E}_{q(x_1|x_0)} \left[\log p_\theta(x_0|x_1) + \log p(x_1) - \log q(x_1|x_0) \right] \\
&= \mathbb{E}_{q(x_1|x_0)} \left[\log p_\theta(x_0|x_1) \right] - \mathrm{KL}\left[q(x_1|x_0) \| p(x_1) \right]
\end{aligned}$$

这个式子分为了两项, 其中

1. 第一项 $\mathbb{E}_{q(x_1|x_0)} \left[\log p_\theta(x_0|x_1) \right]$ 是一个期望值，它表示的是给定潜在变量 x_1 后生成原始图像 x_0 的对数似然。这个期望, 也即其下标 $q(x_1|x_0)$ 表示前向过程, 是一个已知的过程 , 所以我们可以通过采样 x_1 来近似计算这个期望值
2. 第二项 $\mathrm{KL}\left[q(x_1|x_0) \| p(x_1) \right]$ 是是两个已知分布之间的 KL 散度。而经过上面的计算, 我们知道 $q(x_1|x_0)$ 与 $p(x_1)$ 均服从正态分布, 这使得 KL 散度可以解析计算 (两个正态分布之间的 KL 散度可以用公式直接计算).

从而整个变分下界可以计算. 通过最大化这个变分下界的形式, 我们可以最大化似然函数, 找到最适合的 θ 值.

DDPM 中的最大化下界

我们只需要将扩散模型过程整理_Nekasu > 用 VAE (变分自编码器)类比中的 x_1 变成 $x_0 : x_T$ 即可得到 DDPM 的推导过程.

DDPM 的对数似然函数

根据上面的推导, 我们知道 DDPM 有似然函数如下:

$$\log P_\theta(x_0) = \log\left[\int_{x_1:x_T} p(x_T)\cdot p_\theta(x_{T-1}|x_T)\cdot p_\theta(x_{T-2}|x_{T-1})\cdots p_\theta(x_1|x_2)\cdot p_\theta(x_0|x_1)dx_1:x_T\right]$$

DDPM 中引入前向过程

从而有以下推导:

$$\begin{aligned}
\log P_\theta(x_0) &= \log\left[\int_{x_1:x_T} p(x_T)\cdot p_\theta(x_{T-1}|x_T)\cdot p_\theta(x_{T-2}|x_{T-1})\cdots p_\theta(x_1|x_2)\cdot p_\theta(x_0|x_1)dx_1:x_T\right]\\
&= \log\left[\int_{x_1:x_T} p_\theta(x_0,x_1,\cdots,x_T)dx_1:x_T\right]\\
&= \log\left[\int_{x_1:x_T} \frac{\prod_{t=1}^T q(x_t|x_{t-1})}{\prod_{t=1}^T q(x_t|x_{t-1})}p_\theta(x_0,x_1,\cdots,x_T)dx_1:x_T\right]\\
&= \log\left[\int_{x_1:x_T} \prod_{t=1}^T q(x_t|x_{t-1})\cdot\frac{p_\theta(x_0,x_1,\cdots,x_T)}{\prod_{t=1}^T q(x_t|x_{t-1})}dx_1:x_T\right]
\end{aligned}$$

在上面这个推导中的第 3 个等式, 我们仿照 VAE 中的推导, 将被积函数变形为了 1· 被积函数的形式.

- 在 VAE 中, 这个 1 是分子分母均为 $q(x_1|x_0)$ 的分式.
  - q 表示的是前项过程, 也就是从 x_0 获得噪声图像 x_1 的过程
- 在 DDPM 中, 这个 1 是分子分母均为 $\prod_{t=1}^T q(x_t|x_{t-1})$ 的分式
  - 同样的, q 表示的是前项过程, 也就是从 x_0 获得一系列噪声图的过程.
  - 与 VAE 中仅有一个加早过程不同, DDPM 中具有 T 个加噪过程, 即 $q(x_t|x_{t-1}), t=1,2,3,\cdots,T$ 这 T 个加噪过程.
  - 所以需要将这 T 的加噪过程相乘, 也即 $\prod_{t=1}^T q(x_t|x_{t-1})$

对于上面都推导中的累乘 $\prod_{t=1}^T q(x_{t-1}|x_t)$, 我们可以有如下的化简:

$$\prod_{t=1}^T q(x_{t-1}|x_t) = q(x_0|x_1)\cdot q(x_1|x_2)\cdots q(x_{T-1}|x_T) = q(x_1,\cdots,x_T|x_0)$$

- 为了简便表示, 我们将化简后的结果 $q(x_1,\cdots,x_T|x_0)$ 中的 $x_1,x_2,\cdots,x_T$ 记做 $x_{1:T}$,
  - 因此有 $q(x_1,\cdots,x_T|x_0)$ 可以记作 $q(x_{1:T}|x_0)$.
- 类似的, 我们将分式中的分母 $p_\theta(x_0,x_1,\cdots,x_T)$ 记做 $p_\theta(x_{0:T})$.

将这个累乘的化简继续代入公式进行推导, 有

$$\begin{aligned}
\log P_\theta(x_0) &= \log\left[\int_{x_1:x_T} \prod_{t=1}^T q(x_{t-1}|x_t)\cdot\frac{p_\theta(x_{0:T})}{\prod_{t=1}^T q(x_{t-1}|x_t)}dx_1:x_T\right]\\
&= \log\left[\int_{x_1:x_T} q(x_{1:T}|x_0)\cdot\frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)}dx_1:x_T\right]\\
&= \log\mathbb{E}_{q(x_{1:T}|x_0)}\left[\frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)}\right]\\
&\geq \mathbb{E}_{q(x_{1:T}|x_0)}\left[\log\frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)}\right] \qquad [\text{琴生不等式, 如果函数}\varphi\text{为凹函数, 则有}\varphi(\mathbb{E}(x))\geq\mathbb{E}(\varphi(x))]
\end{aligned}$$

DDPM 分离参数θ与常数

实际上, 我们真正想要计算的是包含参数 θ 的部分, 所以我们需要将不含参数 θ 的部分与含有 θ 的部分分开.

对于 $p_\theta(x_{0:T}) = p_\theta(x_0, x_1, \cdots, x_T)$ 而言, $p(x_T)$ 是一个与 θ 无关的量, 因为 x_T 是直接从标准正态分布中获取的. 所以我们将 $p_\theta(x_{0:T})$ 替换为 $p(x_T) \cdot \prod_{t=1}^{T-1} p_\theta(x_{t-1}|x_t)$, 有以下推导

$$
\begin{aligned}
\log P_\theta(x_0) &\geq \mathbb{E}_{q(x_{1:T}|x_0)}\left[\log \frac{p_\theta(x_{0:T})}{q(x_{1:T}|x_0)}\right] \\
&= \mathbb{E}_{q(x_{1:T}|x_0)}\left[\log \frac{p(x_T) \cdot \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t)}{q(x_{1:T}|x_0)}\right] \\
&= \mathbb{E}_{q(x_{1:T}|x_0)}\left[\log p(x_T) + \sum_{t=1}^{T} \log\left[p_\theta(x_{t-1}|x_t)\right] - \log q(x_{1:T}|x_0)\right]
\end{aligned}
$$

此处分解联合分布 $p(x_T) \cdot \prod_{t=1}^{T-1} p_\theta(x_{t-1}|x_t)$ 使我们能够逐步处理每一个时间步的生成过程，有助于将整体问题分解为多个子问题

而 $q(x_{1:T}|x_0)$ 则是表示前向过程, 是一个与 θ 无关的值. 同样的, 我们希望分解这个联合分布以逐步处理每一个时间步的生成过程, 将问题分成多个子问题, 所以可以进行如下替换:

$$
q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}|x_0)
$$

因此有:

$$
\begin{aligned}
\log P_\theta(x_0) &\geq \mathbb{E}_{q(x_{1:T}|x_0)}\left[\log p(x_T) + \sum_{t=1}^{T} \log\left[p_\theta(x_{t-1}|x_t)\right] - \log q(x_{1:T}|x_0)\right] \\
&= \mathbb{E}_{q(x_{1:T}|x_0)}\left[\log p(x_T) + \sum_{t=1}^{T} \log\left[p_\theta(x_{t-1}|x_t)\right] - \sum_{t=1}^{T} \log\left[q(x_t|x_{t-1}|x_0)\right]\right] \\
&= \mathbb{E}_{q(x_{1:T}|x_0)}\left[\log p(x_T) + \sum_{t=1}^{T} \log\left[\frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1}|x_0)}\right]\right] \\
&= \mathbb{E}_{q(x_{1:T}|x_0)}\left[\log p(x_T)\right] + \mathbb{E}_{q(x_{1:T}|x_0)}\left[\sum_{t=1}^{T} \log\left[\frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1}|x_0)}\right]\right] \\
&= \mathbb{E}_{q(x_{1:T}|x_0)}\left[\log p(x_T)\right] + \sum_{t=1}^{T} \mathbb{E}_{q(x_{1:T}|x_0)}\left[\log\left[\frac{p_\theta(x_{t-1}|x_t)}{q(x_t|x_{t-1}|x_0)}\right]\right] \\
&= \mathbb{E}_{q(x_{1:T}|x_0)}\left[\log p(x_T)\right] + \sum_{t=1}^{T} \mathbb{E}_{q(x_{1:T}|x_0)}\left[D_{KL}\left(q(x_t|x_{t-1}) \| p_\theta(x_{t-1}|x_t)\right)\right]
\end{aligned}
$$

这个式子分为了两项, 其中

1. 第一项 $\mathbb{E}_{q(x_{1:T}|x_0)}\left[\log p(x_T)\right]$ 是一个与 θ 无关的值, 在优化时可以忽略
2. 第二项 $\sum_{t=1}^{T} \mathbb{E}_{q(x_{1:T}|x_0)}\left[D_{KL}\left(q(x_t|x_{t-1}) \| p_\theta(x_{t-1}|x_t)\right)\right]$ 是已知分布之间的 KL 散度的和。而经过上面的计算, 我们知道 $q(x_t|x_{t-1})$ 与 $p_\theta(x_{t-1}|x_t)$ 均服从正态分布, 这使得 KL 散度可以解析计算 (两个正态分布之间的 KL 散度可以用公式直接计算).

从而整个公式变得容易计算

DDPM 的总结

实际上, 本人推导的结果与原始论文中推导的结果不太相同.

原始论文中的结果中包含了 3 项, 而本人的推导仅有 2 项. 导致这种区别的原因与说明如下:

1. 原因：原始论文中, 为了能够让其更好运算, 并获得更好的效果, 将 KL 散度进一步细分；而本人并未做这样的工作, 主要是为了降低理解门槛, 并与 VAE 的推导同步.
2. 说明：虽然结果不同, 但理解起来并无区别. 也即, 本人的推导更偏向于理解, 而原始论文的推导更偏向于实践.

极大似然估计（MLE）的核心思想

1. **目标**：
   - 极大似然估计（MLE）的目标是估计模型参数 θ，使得观测数据在给定模型参数下的似然函数 $L(\theta|x)$ 最大化。
2. **直观理解**：
   - 如果似然函数 L 越大，说明观测数据在当前模型参数 θ 下的概率越高，也就是说当前取得的参数 θ 越合理。
3. **数学表达**：
   - 给定数据集 $\{x_i\}_{i=1}^N$，似然函数 $L(\theta|x)$ 表示为观测数据在给定参数 θ 下的联合概率密度函数 $p(x|\theta)$：

$$L(\theta|x) = p(x_1, x_2, \ldots, x_N|\theta) = \prod_{i=1}^{N} p(x_i|\theta)$$

   - 极大似然估计通过最大化对数似然函数来简化计算：

$$\hat{\theta} = \arg\max_\theta L(\theta|x) = \arg\max_\theta \sum_{i=1}^{N} \log p(x_i|\theta)$$

最大化似然函数的动机

- **合理参数的定义**：最大化似然函数 $L(\theta|x)$ 是为了找到一组参数 θ，使得在这些参数下，观测到的数据的概率最大。因此，如果似然函数 L 越大，说明当前取得的参数 θ 越合理，这与极大似然估计的目标完全一致。
- **优化问题**：最大化似然函数实际上是一个优化问题，我们通过寻找使得似然函数最大的参数来实现这一目标。

变分下界（ELBO）和近似计算

由于直接计算似然函数可能非常困难，特别是在高维度和复杂模型中，我们引入变分下界（ELBO）来近似计算。

- **变分下界的目的**：通过引入一个辅助分布q，我们可以将原本难以计算的似然函数转化为变分下界的形式，并通过最小化前向过程q和逆向过程p_θ之间的 KL 散度来进行优化：

$$\text{ELBO} = \mathbb{E}_q\left[\log p_\theta(x_{0:T}) - \log q(x_{0:T}|x_0)\right]$$

1. 输入一个自高斯分布中获取的噪声图像
2. 经过神经网络的操作, 使噪声图像变成一个其他复杂分布的一个实例
3. 这个实例满足的分布与真实分布越接近越好