

Deepia. Youtube DDPM.

DDPM. 2020. 截至25年. 利用年限 20.000t

$$x_0 \xrightarrow{q(x_t|x_0)} x_t$$

$$x_t = x_0 + \sqrt{\beta} \cdot \varepsilon. \quad \varepsilon \sim N(0, I).$$

$$q(x_t|x_0) = N(x_0, \beta I)$$

如果我们继续加噪。

$$x_1 \xrightarrow{q(x_2|x_1)} x_2$$

$$x_2 = x_1 + \sqrt{\beta} \cdot \varepsilon. \quad \varepsilon \sim N(0, I)$$

$$q(x_2|x_1) = N(x_1, \beta I) = N(x_0, 2\beta I)$$

所以一直加噪。最后只是在分布上引入了很大的方差。

而我们希望，加噪候最终结果是 Standard Gaussian.

Standard Normal distribution. $N(0, I)$

如果一直直接加噪. $x_{1000} \sim N(x_0, 1000\beta I)$, Variance Explodes!

方差爆炸。

Objective: $q(x_t|x_0) \xrightarrow{t \rightarrow \infty} N(0, I)$

$$x_{t-1} \xrightarrow{q(x_t|x_{t-1})} x_t$$

$$x_t = \sqrt{1-\beta} \cdot x_{t-1} + \sqrt{\beta} \cdot \varepsilon$$

我们省去微增量等。

$$x_t = \sqrt{\alpha_t} \cdot x_0 + \sqrt{1-\alpha_t} \cdot \varepsilon$$

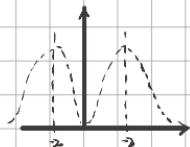
现在先不要去想为什么这能一步完成。

而我们还需要 DPPM. 去扩散。

$$\overline{\alpha_t} = \prod_{i=1}^t (1-\beta_i). \quad \text{在 } t \rightarrow \infty, \sqrt{\overline{\alpha_t}} \rightarrow 0, \\ \sqrt{1-\overline{\alpha_t}} \rightarrow 1.$$

我们很好地完成了 Objective.

Let us Build on Experiment. [www.github.com/Tslaloln/MyDDPM](https://github.com/Tslaloln/MyDDPM)

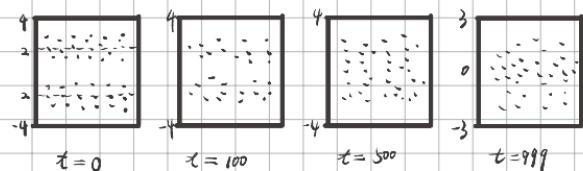


一个双峰混合高斯模型，其分布为

$$\text{Mix}(N(-2, 0.5), N(2, 0.5))$$

训练为 $N = 800$ 个点，拟合而来。

$T_{\text{steps}} = 1000$, 向前扩散 1000 次。



明显双峰分布 \longrightarrow 标准正态分布。



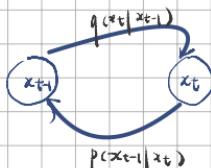
当然，现在我们已知 $x_t = \sqrt{\alpha_t} \cdot x_0 + \sqrt{1-\alpha_t} \cdot \varepsilon$

我们可以令任意时间在 $[0, T]$ 求解 x_t 的分布

很多视频， $q(x_t | x_0)$ —— 就是概率公式。

—— 合成是状态转移公式。很高

现在我们引入 $p(x_{t+1} | x_t)$



$p(x_{t+1} | x_t)$ 与网络参数有关

合成 $p_\theta(x_{t+1} | x_t)$

现在有 θ , 如何训练 θ ? $\Rightarrow \text{Maximum } P_{\theta}(x_0 | x_T)$

等价于 $\text{Minimize } -\log P_{\theta}(x_0 | x_T)$,

稍微翻译一下。你现在有 $x_0 \sim p_{\text{data}}(x)$.

我们 $x_0 \rightarrow x_T$, $x_T \sim N(\mu, I)$.

现在 Model_{θ} , 致于 $\text{Model}_{\theta}(x_T) = x_0$,

我们希望 x_0 不可能是 x_0 . (最大似然)

联合分布.

$$q(x_1=x_T | x_0) = \prod_{i=1}^T q(x_i | x_{i-1})$$

这是前向扩散过程的, 那么反向生成的联合分布?

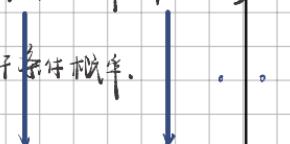
$$p_{\theta}(x_0, x_1, x_2, \dots, x_{T-1}, x_T) = p(x_T) \cdot \prod_{i=1}^{T-1} p_{\theta}(x_{t+1} | x_t)$$

与正向过程不同, 反向过程不带条件.

因为反向是从 $x_T \sim N(\mu, I)$ 进行, 作者认为 x_T 是不带先验知识的

$$q(x_1, x_2, \dots, x_T | x_0) = q(x_1 | x_0) \cdot q(x_2 | x_1, x_0) \cdot q(x_3 | x_2, x_1, x_0) \cdots q(x_T | x_{T-1}, \dots, x_0)$$

贝叶斯定理, 将联合概率分解为若干条件概率.



由于每一步都独立.

马尔可夫过程.

完全一致了

$$q(x_1, x_2, \dots, x_T | x_0) = q(x_1 | x_0) \cdot q(x_2 | x_1) \cdot q(x_3 | x_2) \cdots q(x_T | x_{T-1})$$

| 简写

$$q(x_1=x_T) = q(x_1 | x_0) + q(x_2 | x_1) + \dots + q(x_T | x_{T-1})$$

这个其实很理解, 类似于

$$P(A, B, C) = P(A) \cdot P(B | A) \cdot P(C | B)$$

正向联合分布概率密度。 $p(x_1 \dots x_T | x_0) = \prod_{i=1}^T p(x_i | x_{i-1})$

反向联合分布概率密度。 $p_\theta(x_0 \dots x_T) = p(x_T) \cdot \prod_{i=1}^T p_\theta(x_{i-1} | x_i)$

暂时先放着， p_θ 有用，因为 Maximum Likelihood Estimate 要求：

$$\arg\max_\theta \prod_{i=1}^N p_\theta(x_i) = \arg\min_\theta -\sum_{i=1}^N \log p_\theta(x_i)$$

这里的 $x_i, i=1, 2, \dots, N, x_i \sim p_{\text{data}}(x)$ ，与 x_0 截然不同呀。

我们现在就假设 p_{data} 上只有一张图片 x_0 ，上式就改为：

$$\arg\min_\theta -\log p_\theta(x_0),$$

那么这玩意 $p_\theta(x_0)$ ，还是处理不了。但是！我们有联合分布

而 $p_\theta(x_0)$ 是 $p_\theta(x_0 \dots x_T)$ 的边缘分布！

$$p_\theta(x_0) = \int_{x_1} \left[\int_{x_2} \left[\int_{x_3} \left[\int_{x_T} p_\theta(x_0 \dots x_T) \cdot dx_T \right] \cdot dx_{T-1} \right] \dots dx_1 \right]$$

$$= \int_{x_1=x_T} p_\theta(x_0 \dots x_T) \cdot dx_1 = dx_T$$

$$p_\theta(x_0) = \int_{x_1=x_T} p_\theta(x_0 \dots x_T) \cdot dx_1 = dx_T$$

这个部分的含义是，计算 Model 生成 x_0 的概率。

就从 x_T 起，将所有 $x_T \rightarrow x_{T-1} \rightarrow \dots \rightarrow x_1 \rightarrow x_0$ 路径全部求取出来，求其概率之和。

这很明显是无法做到的，但是我们不需要去真把 $p_\theta(x_0)$ 算出来，我们实际上想做的只是

Minimize $-\log p_\theta(x_0)$ 而已！

那么处理这个东西我们还是做得到的。

$$-\log P_{\theta}(x_0) = -\log \int_{x_1=x_1} p_{\theta}(x_0=x_1) dx_1$$

之前的前向联合概率分布

修正上?

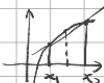
$$\text{上下同时乘除. } = -\log \int_{x_1=x_1} q(x_1=x_1|x_0) \cdot \frac{p_{\theta}(x_0=x_1)}{q(x_1=x_1|x_0)} dx_1$$

$$= -\log E_{q(x_1=x_1|x_0)} \left[\frac{p_{\theta}(x_0=x_1)}{q(x_1=x_1|x_0)} \right]$$

$$-E_{q(x_1=x_1|x_0)} \left[\log \frac{p_{\theta}(x_0=x_1)}{q(x_1=x_1|x_0)} \right]$$

这里是因为

根据著名的詹森不等式



$$\log \left(\frac{x_1+x_2}{2} \right) \geq \frac{1}{2} (\log x_1 + \log x_2)$$

$$\log E(x) \geq E(\log(x))$$

$$\int_{x_1=x_1} q(x_1=x_1|x_0) \cdot dx_1$$

$$x_1=x_1$$

$$= \left(\int_{x_1} q(x_1|x_0) \cdot dx_1 \right) \cdot \left(\int_{x_2} q(x_2|x_1) \cdot dx_1 \right) \cdots \left(\int_{x_T} q(x_T|x_{T-1}) \cdot dx_1 \right)$$

$$= I \times I \times \cdots \times I$$

$$= 1.$$

那现在，我们从 Minimize $-\log P_{\theta}(x_0)$

$$\rightarrow \text{Minimize } -E_{q(x_1=x_1|x_0)} \left[\log \frac{p_{\theta}(x_0=x_1)}{q(x_1=x_1|x_0)} \right]$$

或者说 Maximum $\log P_{\theta}(x_0)$

$$\rightarrow \text{Maximum } E_{q(x_1=x_1|x_0)} \left[\log \frac{p_{\theta}(x_0=x_1)}{q(x_1=x_1|x_0)} \right]$$

希望 $\log P_{\theta}(x_0)$ 尽量大，
那 $\log P_{\theta}(x_0)$ 的下界尽量大。

下界. Lower Bound.

(Evidence Lower Bound, ELBO)

$$\text{我们现在继续 Minimize } -E_{q(x_t|x_0)} \left[\log \frac{p_\theta(x_0|x_t)}{q(x_t|x_0|x_0)} \right]$$

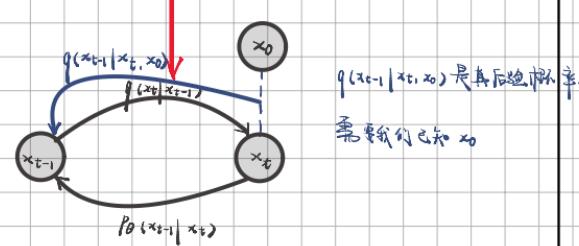
代数运算 贝叶斯公式

$$-E_q \left[D_{KL}(q(x_t|x_0) \| p(x_t)) + \sum_{t>1} D_{KL}(q(x_{t-1}|x_0, x_0) \| p_\theta(x_{t-1}|x_t)) - \log p_\theta(x_0|x_1) \right]$$

$$-E_q \left[D_{KL}(q(x_t|x_0) \| p(x_t)) + \sum_{t>1} D_{KL}(q(x_{t-1}|x_0, x_0) \| p_\theta(x_{t-1}|x_t)) - \log p_\theta(x_0|x_1) \right]$$

鉴于我们是做了 $\arg\min_\theta$,

而这项不含 θ



$q(x_{t-1}|x_t, x_0)$ 是真后验概率.

需要我们已知 x_0

这也这操作听怪

$-\log p_\theta(x_0|x_1)$ 尽量小，即 $p_\theta(x_0|x_1)$ 尽量大。

即给你带噪图像 x_1 ，你最好给我的 $Model_\theta(x_1) \rightarrow x_0$

那么我们知道 diffusion model 的效果其实是很好的，所以反向去噪到 x_1 ， x_1 已经非常接近真实分布 x_0 ，所以通常我们说可以忽略这一项。

关于没听懂那点，那就取 $q(x_{t-1} | x_t, x_0) \parallel p_0(x_{t-1} | x_t)$ 。
如何操作一通到两个 Gaussian distribution.

我们再回到前面扩散过程：

$$x_t = \sqrt{dt} \cdot x_0 + \sqrt{1-dt} \cdot \varepsilon, \text{ 这是我们实际用的.}$$

$$\text{但 } x_t = \sqrt{dt} \cdot x_{t-1} + \sqrt{1-dt} \cdot \varepsilon, \text{ 也没问题.}$$

$$\text{其中 } \varepsilon \sim N(0, I)$$

确定项.

随机项.

高斯分布的线性变换性质后。

$$\text{若 } \varepsilon \sim N(0, I). \text{ 又 } A\varepsilon + b \sim N(b, AA^T)$$

常数.

$$\text{于是 } x_t = \sqrt{1-dt} \cdot \varepsilon + \sqrt{dt} \cdot x_{t-1} \sim N(\sqrt{dt} \cdot x_{t-1}, (1-dt)I), \text{ 此时认为 } x_{t-1} \text{ 是常数.}$$

$$\text{即其形式是 } x_t | x_{t-1} \sim N(\sqrt{dt} \cdot x_{t-1}, (1-dt)I).$$

举个例子. $X \sim \text{Uniform}(-1, 1)$, $\varepsilon \sim N(0, 1)$, $Y = X + \varepsilon$.

那 $Y | X \sim N(X, 1)$

我们现在能接受 $x_t | x_{t-1} \sim N(\sqrt{dt} \cdot x_{t-1}, (1-dt)I)$ 这一写法.

$$\text{那 } u \text{ 为 } x_t = \sqrt{dt} \cdot x_0 + \sqrt{1-dt} \cdot \varepsilon, \text{ 有.}$$

$$x_t \text{ 还可以写成 } x_t | x_0 \sim N(\sqrt{dt} \cdot x_0, (1-dt)I),$$

$$t \rightarrow \infty, \text{ 则 } \sqrt{dt} \rightarrow 0, x_t | x_0 \rightarrow N(0, I).$$

$$\begin{aligned} & \downarrow \\ & N(0, I) \text{ 不含 } x_0 \\ & x_t \rightarrow N(0, I) \end{aligned}$$

回到 $q(x_{t-1} | x_t, x_0)$, 使用贝叶斯公式.

$$\begin{aligned} q(x_{t-1} | x_t, x_0) &= \frac{q(x_t | x_{t-1}, x_0) \cdot q(x_{t-1} | x_0)}{q(x_t | x_0)} \\ &\stackrel{\text{Markov Process}}{=} \frac{(q(x_t | x_{t-1}) \cdot q(x_{t-1} | x_0))}{q(x_t | x_0)} \\ &= \frac{\text{高斯分布} \times \text{高斯分布}}{\text{高斯分布}} \end{aligned}$$

$$q(x_{t-1} | x_t, x_0) \propto q(x_t | x_{t-1}) \cdot q(x_{t-1} | x_0)$$

\rightarrow 起消去作用, 省略即可.

$$q(x_{t-1} | x_t, x_0) \propto q(x_t | x_{t-1}) \cdot q(x_{t-1} | x_0)$$

其中, $x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1-\alpha_t} \cdot \varepsilon$.

$$x_{t-1} = \sqrt{\alpha_{t-1}} \cdot x_0 + \sqrt{1-\alpha_{t-1}} \cdot \varepsilon$$

↓
很复杂的推导后.

$$q(x_{t-1} | x_t, x_0) = N(\hat{\mu}_t(x_t, x_0), \hat{\beta}_t \cdot I)$$

$$\text{其中 } \hat{\mu}_t(x_t, x_0) = \frac{\sqrt{\alpha_t} \cdot (1 - \hat{\beta}_{t-1})}{1 - \hat{\beta}_t} \cdot x_t + \frac{\sqrt{\alpha_{t-1}} \cdot \hat{\mu}_t}{1 - \hat{\beta}_t} \cdot x_0$$

$$\hat{\beta}_t = \frac{1 - \hat{\beta}_{t-1}}{1 - \hat{\beta}_t} \cdot \beta_t$$

我们也得到“高斯 * 高斯 = 高斯”这一结论.

现在我们看 $D_{KL}(q(x_{t-1} | x_t, x_0) || p_0(x_{t-1} | x_t))$

这一边

在 $p_0(x_{t-1} | x_t)$, 我们并没有逆推式!!!

我们无法人为去设定. 如像 forward 时, $x_t = \sqrt{\alpha_t} \cdot x_{t-1} + \sqrt{1-\alpha_t} \cdot \varepsilon$

$$\text{所以 反向去噪时 } x_{t-1} \text{ 就变成 } x_{t-1} = \frac{x_t}{\sqrt{\alpha_t}} - \frac{\sqrt{1-\alpha_t}}{\sqrt{\alpha_t}} \cdot \varepsilon$$

没那么简单啦.

$$\text{我们已知 } q(x_{t-1} | x_0, x_0) = N(\hat{\mu}_t(x_t, x_0), \hat{\beta}_t \cdot I)$$

我们假设 $p_0(x_{t-1} | x_0)$ 也会是 $N(\mu_0, \sigma_0)$ 的样子.

这个假设并不过分.

$$\text{于是 } p_0(x_{t-1} | x_t) = N(\mu_0(x_t), \sigma_0)$$

$$= N(\mu_0(x_t, t), \hat{\beta}_t \cdot I)$$

实际上, 它是
一个已知常数, 它
不需要再测

$$\text{由 } \arg \min D_{KL} \rightarrow 0, R_1: \mu_0(x_t, t) \xrightarrow{\text{应该}} \hat{\mu}_t(x_t, x_0)$$

$$\text{若 } \mu_0(x_t, t) = \frac{\sqrt{\alpha_t} \cdot (1 - \hat{\beta}_{t-1})}{1 - \hat{\beta}_t} \cdot x_t + \frac{\sqrt{\alpha_{t-1}} \cdot \hat{\mu}_t}{1 - \hat{\beta}_t} \cdot x_0$$

$$\mu_\theta(x_t, t) = \tilde{\mu}_t(x_t, x_0)$$

$$= \frac{\sqrt{dt} \cdot (1 - \bar{d}_{t-1})}{1 - \bar{d}_t} \cdot x_t + \frac{\sqrt{dt_{t-1}} \cdot f_t}{1 - \bar{d}_t}$$

在反向过程，我们并不知道 x_0 ，暂用

$$x_t = \sqrt{dt} \cdot x_0 + \sqrt{1-dt} \cdot \varepsilon \Rightarrow x_0 = \frac{1}{\sqrt{dt}} (x_t - \sqrt{1-dt} \cdot \varepsilon)$$

$$= \frac{1}{\sqrt{dt}} \left(x_t - \frac{\beta_t}{\sqrt{1-\bar{d}_t}} \cdot (\varepsilon) \right)$$

代换掉.

代换后这玩意还是不知道.

构建 $\varepsilon_\theta(x_t, t)$ ，去学习【与 x_0, x_t 相关的】 ε （是这个 ε 导致了 $x_t = \sqrt{dt} \cdot x_0 + \sqrt{1-dt} \cdot \varepsilon$ ，今天反向去噪）

于是损失函数变为 $MSE(\varepsilon_\theta(x_t, t), \varepsilon)$

如果 ε_θ 学到了什么样的 ε 能把 x_t 还原回 x_0 ，
那就学会了 x_0 的分布）

现在，我们有了损失函数 $MSE(\varepsilon_{\theta}(x_t, t), \varepsilon)$ 。

并经过艰苦的训练，实现了 ε_{θ} 的收敛。

那我们能不能直接在 $t=1000$ 时。

$$由 \quad x_0 = \frac{1}{\sqrt{\alpha_t}}(x_t - \sqrt{1-\alpha_t} \cdot \varepsilon), \text{ 恢复 } x_0?$$

那这可能不太行，对吧，毕竟去噪时其实是一步一步做的。

OK，那我们还是一步步地去噪。

现在 $p_{\theta}(x_{t-1} | x_t) = N(\mu_{\theta}(x_t, t), \beta_t \cdot I)$

$\mu_{\theta}(x_t, t)$ 已知， $\mu_{\theta}(x_t, t) = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\alpha_t}} \cdot \varepsilon_{\theta})$

$$x_{t-1} = \mu_{\theta}(x_t, t) + \sqrt{\beta_t} \cdot z$$

$$= \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{\beta_t}{\sqrt{1-\alpha_t}} \cdot \varepsilon_{\theta}(x_t, t) \right) + \sqrt{\beta_t} \cdot z$$

其中 $z \sim N(0, I)$ 至此我们才获取了我们反向去噪的递推式。

关于为什么“ $+ \sqrt{\beta_t} \cdot z$ ”

最直接（对我而言）的解释是：

$$p_{\theta}(x_{t-1} | x_t) = N(\mu_{\theta}(x_t, t), \beta_t \cdot I) = N(\mu_{\theta}(x_t, x_0), \beta_t \cdot I)$$

即最大似然的内存要求。

这非常明显，也解释了 ε 是高斯分布的原因。

而另一方面，正向过程 $x_t (t=1 \dots T)$ 先一个高斯分布。

即给定 $x_0, x_t (t=1 \dots T)$ 都是确定的， $q(x_t | x_0)$ 都是高斯分布。

那么从反向过程与正向过程朴素的对称性出发。

现在反向去噪时， x_T 是给定的，而 (x_0, T) 本身也是确定的。

如果你不引入 $z \sim N(0, I)$ ，反向过程在 x_T 给定时， $x_t (t=T-1 \dots 0)$

全部是确定的，不会有 $p(x_0, x_T)$ 是高斯分布。

那这当然谈不上 $D_{KL}[q(x_{t-1} | x_t, x_0) || p_{\theta}(x_{t-1} | x_t)] \rightarrow 0$ 。

我们的正/反向过程也不对称了。

缺个总结.

但是这玩意儿真的靠人类想出来的吗?

太黑了. 告别现在才勉强明白.