DDPM

理论部分

主要参考资料：[生成扩散模型漫谈（二）：DDPM = 自回归式VAE](), [Diffusion Model 入门（1）——概率生成式模型概述]()

从VAE概率角度看

变分：在真实后验概率较为复杂时，使用一个近似概率分布去近似真是概率分布的过程就叫做变分

推理：给定的隐变量时，用似然度对被观测变量进行采样。

扩散生成模型，本质上仍然是一个最大似然概率模型，也就是最大化解码器（生成器）在这个参数下，输出这个样本x的概率$p_\theta(x)$其中θ指的是生成器参数，x指的是数据样本。

Loss推导

VAE回顾

首先让我们回归VAE的Loss部分，对于VAE而言，我们试图学习一种映射关系，可以将输入图像 x 映射到隐空间 z，然后再从隐空间 z 重构出原始图像，因此，我们想到了似然，这里的似然需要跟模型的隐变量进行挂钩，所以我们将似然写作边缘概率形式$p_\theta(x)=\int p_\theta(x,z)\,dz=\int p_\theta(x|z)p(z)\,dz$，由于神经网络采用的是端到端的学习，所以也就是输入x输出x，所以无法知道每个图像x和隐变量z之间的关系，无法监督的实现这个事情，所以直接计算这个积分是不可行的，因此我们想到了贝叶斯公式，可以有$p_\theta(x)=\frac{p_\theta(x,z)}{p_\theta(z|x)}$，这样我们可以引入一个中间的过程$p_\theta(z|x)$这样每个输入的$x$都能知道其对应的隐变量$z$。然而由于我们$\theta$是解码器，其无法进行计算得到隐变量$z$，所以我们这里需要引入另外一个网络也就是编码器的近似分布$q_\phi(z|x)$用于近似解码器的真实分布$p_\theta(z|x)$。因此我们可以进行公示的推导了

$$p_\theta(x)=\int p_\theta(x,z)\,dz=\int\frac{p_\theta(x,z)}{q_\phi(z|x)}q_\phi(z|x)dz=\mathbb{E}_{q_\phi(z|x)}\frac{p_\theta(x,z)}{q_\phi(z|x)}$$

对于解码器而言，他是能够知道$p_\theta(x|z)$的过程的，因此我们可以接着直觉上的对公式进行进一步的化简。

$$\mathbb{E}_{q_\phi(z|x)}\frac{p_\theta(x,z)}{q_\phi(z|x)}=\mathbb{E}_{q_\phi(z|x)}\frac{p_\theta(x|z)p_\theta(x)}{q_\phi(z|x)}$$

我们一般习惯性求的是对数似然，因此有

$$logp_\theta(x)=log\mathbb{E}_{q_\phi(z|x)}\frac{p_\theta(x|z)p_\theta(x)}{q_\phi(z|x)}$$

利用詹森不等式，我们把期望和对数进行交换有

$$logp_\theta(x)=log\mathbb{E}_{q_\phi(z|x}\frac{p_\theta(x|z)p_\theta(z)}{q_\phi(z|x)}\geq\mathbb{E}_{q_\phi(z|x)}log\frac{p_\theta(x|z)p_\theta(z)}{q_\phi(z|x)}$$

因此就求得了变分下界，我们要最大化变分下界也就是最小化负变分下界。

$$\mathcal{L}=-\mathbb{E}_{q_\phi(z|x)}log\frac{p_\theta(x|z)p_\theta(z)}{q_\phi(z|x)}=D_{KL}(q_\phi(z|x)||p_\theta(z))+\mathbb{E}_{q_\phi(z|x)}-logp_\theta(x|z)$$

这样，我们就最终从似然的角度推导出了VAE的Loss。

其中第一项是先验分布和编码器近似分布的KL散度，目的是让我们模型的潜在空间分布近似于先验假设的标准正态分布，第二项是重构损失，也就是让我给定的x，通过这个x得到的z，通过这个z恢复的数据x之间的差异最小化。实际上VAE是一个对抗的过程，一方面，我们希望模型能够尽可能好地重构输入数据（通过最小化重构损失），这样会让隐空间变得复杂，不接近于先验的标准正态分布；另一方面，我们也希望模型的潜在空间分布能够接近我们的先验假设（通过最小化KL散度损失）。

对于第一项KL散度的公式和第二项为什么是MSE均有推导，这里不在赘述。

DDPM部分

现在正是进入DDPM部分，首先说个定义，DDPM本质上就是VAE的多步版本。所以他仍然是一个最大似然概率模型。因此损失函数自然从似然的角度出发。至于为什么ddpm是多步VAE的版本。主要原因在于。VAE做的工作时直接将一个真实数据分布$p(x)$拟合为一个正态分布$q_\phi(z|x)$，中间的跨度很大，所以效果不会特别的好。因此我们可以考虑做一个渐变的工作，将真实分布变为正态分布的过程，变为一个多步过程，每一步变味正态分布的变化都是极小的，模型对于这种极小的差异学习能力会非常的强，从而让生成的效果变得极佳。

首先我们从马尔科夫性质开始说起

$p_\theta(x_0,x_1,x_2,...,x_T)$

$= p_\theta(x_0,x_1,x_2,...,x_{T-1}|x_T)p_\theta(x_T) $

$= p_\theta(x_0,x_1,x_2,...,x_{T-2}|x_{T-1},x_T)p(x_{T-1}|x_T)p_\theta(x_T)$

$= ...$

$= p_\theta(x_0|,x_1,x_2,...,x_{T-2},x_{T-1},x_T)p_\theta(x_1|x_2,...,x_{T-1},x_T)...p(x_{T-1}|x_T)p_\theta(x_T)$

但是，由马尔科夫性质，我们可以得知，当前状态仅与上一时刻状态有关。这个形状虽然表示x_{i+1}只和x_i的性质有关，但是并不代表x_{i+1}不受x_{i-1}影响，从而x_{i+1}与x_{i-1}是独立的，而是x_{i-1}对x_{i+1}的所有影响都由x_i传递给x_{i+1}

因此上述公式可以化简为

$= p_\theta(x_0|,x_1,x_2,...,x_{T-2},x_{T-1},x_T)p_\theta(x_1|x_2,...,x_{T-1},x_T)...p(x_{T-1}|x_T)p_\theta(x_T)$

=$p_\theta(x_T)\prod_{t\geq 1}p_\theta(x_{t-1}|x_t) \quad Eq(1)$

同理对$q_\phi(x_1,x_2,...,x_T|x_0)$可以化简为

$= \prod_{t\geq 1}q_\phi(x_t|x_{t-1}) \quad Eq(2)$

既然是最大似然概率模型，仍然从极大似然估计开始。为了省略步骤，我们直接从VAE的VLB开始改造。也就是说

$-logp_\theta(x) \leq \mathbb{E}_{q_\phi(z|x)}[-log\frac{p_\theta(x,z)}{q_\phi(z|x)}]$

既然是多步VAE，那么我们就假设隐变量的分布有多个，因此$z=z1,z2,...,zt$为了与论文统一公式，我们令$x=x0,zt=xt,t\in\{1,2,...,t\}$有

$-logp_\theta(x) \leq \mathbb{E}_{q_\phi(z|x)}[-log\frac{p_\theta(x,z)}{q_\phi(z|x)}] = \mathbb{E}_{q_\phi(x_1,x_2,...,x_T|x_0)}[-log\frac{p_\theta(x_0,x_1,x_2,...,x_T)}{q_\phi(x_1,x_2,...,x_T|x_0)}]$

由于我们在算loss的时候，实际上是多个x_0的一起输入，然后对整个数据集的每个样本的loss取平均因此公式又可以写作

$\mathbb{E}_{q_\phi(x_1,x_2,...,x_T|x_0)}[-log\frac{p_\theta(x_0,x_1,x_2,...,x_T)}{q_\phi(x_1,x_2,...,x_T|x_0)}] = \mathbb{E}_{q_\phi(x_0)}\mathbb{E}_{q_\phi(x_1,x_2,...,x_T|x_0)}[-log\frac{p_\theta(x_0,x_1,x_2,...,x_T)}{q_\phi(x_1,x_2,...,x_T|x_0)}]$

$= \mathbb{E}_{q_\phi(x_0,x_1,x_2,...,x_T)}[-log\frac{p_\theta(x_0,x_1,x_2,...,x_T)}{q_\phi(x_1,x_2,...,x_T|x_0)}]$

由上面的马尔科夫性质的结论$Eq(1)$和$Eq(2)$有

$= \mathbb{E}_{q_\phi(x_0,x_1,x_2,...,x_T)}[-log\frac{p_\theta(x_T)\prod_{t\ge1}p_\theta(x_{t-1}|x_t)}{\prod_{t>1}q_\phi(x_t|x_{t-1})}]$

$= \mathbb{E}_{q_\phi(x_0,x_1,x_2,...,x_T)}[-log\sum_{t\ge1}\frac{p_\theta(x_{t-1}|x_t)}{q_\phi(x_t|x_{t-1})} - logp_\theta(x_T)] \quad Eq(3)$

接着为了对齐原论文的公式，我这里省略$Eq(3)$中θ和ϕ的表达有

$Eq(3)=\mathbb{E}_{q_\phi(x_0,x_1,x_2,...,x_T)}[-log\sum_{t\ge1}\frac{p_\theta(x_{t-1}|x_{t})}{q_\phi(x_t|x_{t-1})}-logp_\theta(x_T)] $

$= \mathbb{E}_{q(x_0,x_1,x_2,...,x_T)}[-log\sum_{t\ge1}\frac{p(x_{t-1}|x_t)}{q(x_t|x_{t-1})} - logp(x_T)]$

因为当t=1的时候，会与初始条件x_0有关系，所以我把上述式子当t=1的时候拆开有

$= \mathbb{E}_{q(x_0,x_1,x_2,...,x_T)}[-log\sum_{t>1}\frac{p(x_{t-1}|x_t)}{q(x_t|x_{t-1})} - logp(x_T) - \frac{p(x_0|x1)}{q(x_1|x_0)}]$