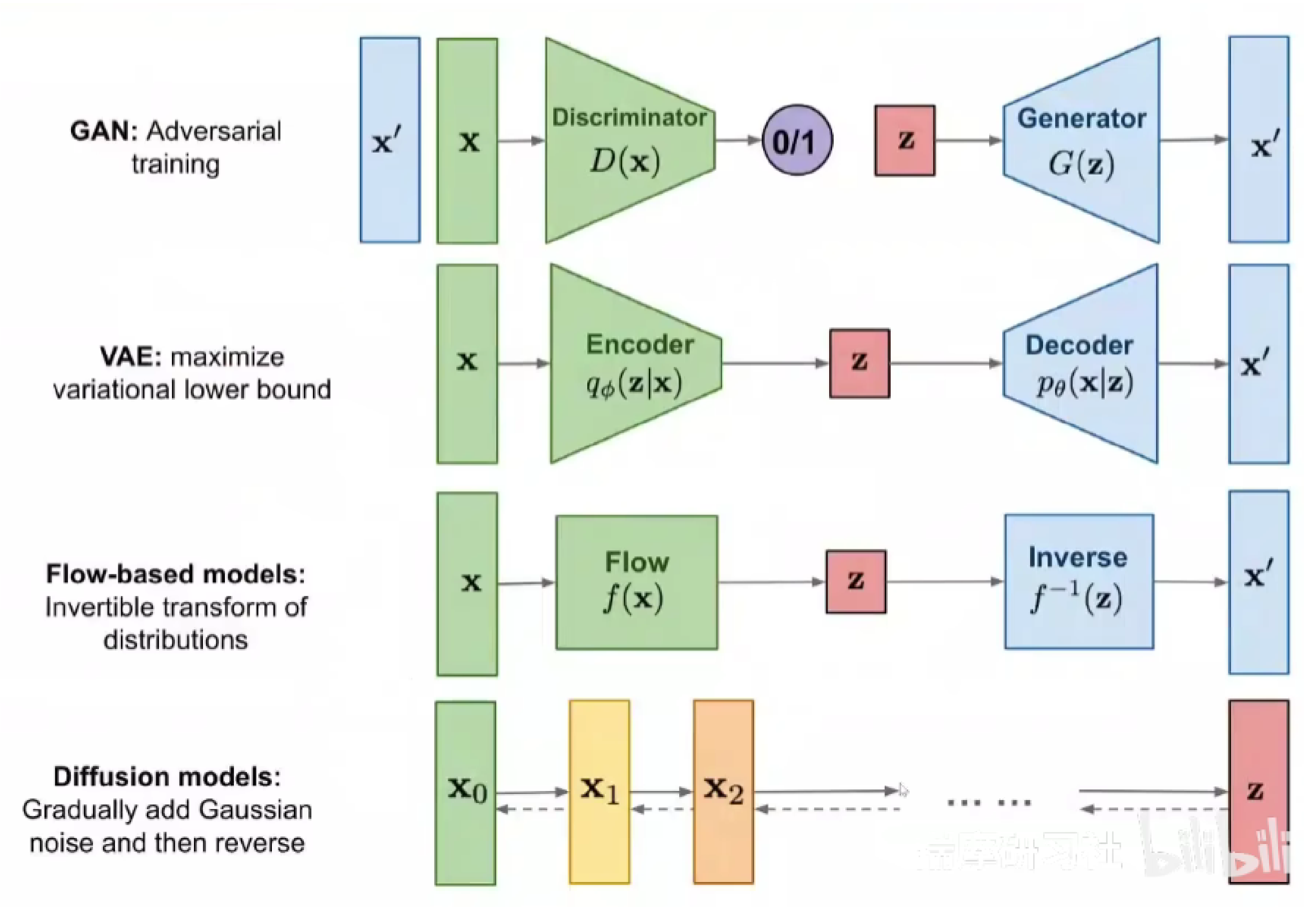
vae

author: doomx
time: 2023.9.23

生成模型的研究背景(fig1)



a. GAN model由于对抗性训练的性质，有潜在的不稳定的训练和较低的多样性。
b. VAE 依赖于代用损失，需要 maximize ELBO (variation lower bound)，不是直接去做p(x)，而是用一个东西去代替。
c. 流模型(Flow-based model) 必须使用专门的架构来构建逆变换

对于许多模态，我们可以将我们观察到的数据视为由**相关的**看不见的**隐变量**表示或生成，我们用**随机变量**z表示。表达这个想法的最佳直觉是通过柏拉图的洞穴寓言。在这个寓言中，一群人一生都被锁在一个山洞里，只能看到投射在他们面前的墙上的二维阴影，这是由看不见的三维物体在火前经过而产生的。对于这样的人来说，他们所观察到的一切，实际上都是由他们永远看不到的更高维度的抽象概念决定的。

ELBO-Evidence Lower Bound

将隐变量z和我们观察到的数据x，用联合分布p(x, z)建模。用最大似然“likelihood-based”的方法，学习一个模型以最大化所有观察到的x的likelihood p(x)。接下来，由两种方法去重构数据

p(x)。

显式地边缘化隐变量z，我们不关心z，我们用一个积分把它积掉，得到我们所关心的p(x):
eq.1

$$p(x) = \int p(x,\ z)\, dz$$

或者，我们可以用概率链法则:
eq.2

$$p(x) = \frac{p(x,\ z)}{p(z \mid x)} = \frac{p(x \mid z)\, p(z)}{p(z \mid x)}$$

上面两个公式，我们想去直接算p(x)是做不到的，首先z的分布我们是很难去处理得到的，所以接下来的思路是用一个代用的损失去替代，也就是提出一个东西，叫ELBO(Evidence Lower Bound)，顾名思义，他是**证据**的下界。

这种情况下，evidence被量化为x的log likelihood，这样，maximize ELBO就成为了优化隐变量模型的代理任务。在最好的情况下，当ELBO被参数化且完美优化时，它与证据完全等价。

我们定义一个变分分布近似，$q_\theta(z \mid x)$是一个**灵活的**变分分布近似， θ 是我们寻求优化的参数，我们通过调整 θ 来增加下限以最大化ELBO。通过这种方式，用**一个变分分布近似去替换隐变量z的分布**，从而获得可用于对真实数据分布p(x)进行建模并从中采样的模型，从而学习 θ 。
eq.3

$$\log p(x) = \int p(x,z)\, dz \quad (\text{Apply eq.1})$$

eq.4

$$= log \int \frac{p(x,\ z)\, q_\theta(z \mid x)}{q_\theta(z \mid x)}\, dz \quad (Multiply\ by\ 1 = \frac{q_\theta(z \mid x)}{q_\theta(z \mid x)})$$

eq.5

$$= \log\ E_{q_\theta(z \mid x)}[\frac{p(x,\ z)}{q_\theta(z \mid x)}] \quad (Definition\ of\ Expectation)$$

eq.6

$$\geq\ E_{q_\theta(z \mid x)}\, [\log \frac{p(x,\ z)}{q_\theta(z \mid x)}] \quad (Apply\ Jensen's\ Inequality)$$

Jensen's Inequality(fig2&3)

凸函数

凸函数是一个定义在某个向量空间的凸子集 C（区间）上的实值函数 f，如果在其定义域 C 上的任意两点 x_1, x_2 ， $0 \leq t \leq 1$ ，有

$$tf(x_1) + (1-t)f(x_2) \geq f\left(tx_1 + (1-t)x_2\right) \qquad (1)$$

也就是说凸函数任意两点的割线位于函数图形上方， **这也是Jensen不等式的两点形式。**

Jensen不等式

若对于任意点集 $\{x_i\}$，若 $\lambda_i \geq 0$ 且 $\sum_i \lambda_i = 1$ ，使用**数学归纳法**，可以证明凸函数 f (x) 满足：

$$f(\sum_{i=1}^{M} \lambda_i x_i) \leq \sum_{i=1}^{M} \lambda_i f(x_i) \qquad (2)$$

公式(2)被称为 Jensen 不等式，**它是式(1)的泛化形式。**

在概率论中，如果把 λ_i 看成取值为 x_i 的离散变量 x 的概率分布，那么公式(2)就可以写成

$$f(E[x]) \leq E[f(x)]$$

其中, $E[\cdot]$ 表示期望。

对于连续变量，Jensen不等式给出了积分的凸函数值和凸函数的积分值间的关系：

$$f(\int xp(x)dx) \leq \int f(x)p(x)dx$$

用了Jensen's Inequality，从等式变到了不等式，需要找到缺失的东西。

现在我们尝试用eq.2去推导ELBO:
eq.7

$$log\, p(x) = log\, p(x) \int q_\theta(z \mid x)\, dz \quad (Multiply\, by\, 1\ =\ \int q_\theta(z \mid x)\, dz)$$

eq.8

$$= \int q_\theta(z \mid x)(log\, p(x))\, dz \quad (Bring\ evidence\ into\ integral)$$

eq.9

$$= E_{q_\theta(z \mid x)}\left[log\, p(x)\right] \quad (Definition\ of\ Expectation)$$

eq.10

$$= E_{q_\theta(z \mid x)}\left[log\, \frac{p(x,\ z)}{p(z \mid x)}\right] \quad (Apply\ equation\ 2)$$

eq.11

$$= E_{q_\theta(z \mid x)}\left[log\, \frac{p(x,\ z)\ q_\theta(z \mid x)}{p(z \mid x)\ q_\theta(z \mid x)}\right] \quad (Multiply\ by\ 1 = \frac{q_\theta(z \mid x)}{q_\theta(z \mid x)})$$

eq.12

$$= E_{q_\theta(z \mid x)}\left[log\, \frac{p(x,\ z)}{q_\theta(z \mid x)}\right] + E_{q_\theta(z \mid x)}\left[log\, \frac{q_\theta(z \mid x)}{p(z \mid x)}\right] \quad (Split\ the\ Expectation)$$

eq.13

$$= E_{q_\theta(z \mid x)}\left[log\, \frac{p(x,\ z)}{q_\theta(z \mid x)}\right] + D_{KL}(q_\theta(z \mid x)\ ||\ p(z \mid x)) \quad (Definition\ of\ KL\ Divergence)$$

eq.14: log p(x) ≥ ELBO

$$\geq\ E_{q_\theta(z \mid x)}\left[log\, \frac{p(x,\ z)}{q_\theta(z \mid x)}\right] \quad (KL\ Divergence\ always\ \geq\ 0)$$

我们这时候也可以得知，用了Jensen's Inequation缺失的就是KL Divergence。

这时可以得知，ELBO项为下界的原因是，evidence与ELBO之间的差异是一个严格非负的KL Divergence，因此ELBO的值永远不会超过evidence。

我们想要优化我们的变分后验（变分分布近似）q_θ(z | x)的参数以逼近真实后验分布p(z | x)。log p(x)为常数，因此ELBO和KL Divergence的和为常数，使ELBO项变大等于使KL Divergence项变小，ELBO可以作为近似学习后验分布的代理任务。

我们需要优化一系列由 θ 参数化的后验分布（变分近似分布）q_θ(z | x)，他被称为自编码器（auto-encoder）。接下来我们进一步解析ELBO：

eq.15

$$E_{q_\theta(z \mid x)}\left[log\, \frac{p(x,\ z)}{q_\theta(z \mid x)}\right] = E_{q_\theta(z \mid x)}\left[log\, \frac{p_\phi(x \mid z)\ p(z)}{q_\theta(z \mid x)}\right] \quad (Chain\ Rule\ of\ Probability)$$

eq.16

$$= E_{q_\theta(z \mid x)}\left[log\, p_\phi(x \mid z)\right] + E_{q_\theta(z \mid x)}\left[log\, \frac{p(z)}{q_\theta(z \mid x)}\right] \quad (Spilt\ the\ Expectation)$$

eq.17

$$= E_{q_\theta(z \mid x)}\left[log\, p_\phi(x \mid z)\right] - D_{KL}(q_\theta(z \mid x)\ ||\ p(z)) \quad (Definition\ of\ KL\ Divergence)$$

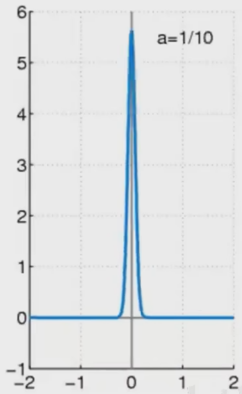
在这种情况下，我们学习了一个变分近似分布 q_θ(z | x)，它被视为**编码器**，可以将输入转换为

可能的隐变量的分布。同时，我们学习了一个**确定性函数** $p_\varphi(x \mid z)$，将给定的隐变量z转换为观测值x，可以将其当作**解码器。**

狄拉克 δ 函数(fig4)

数学中，狄拉克δ函数或简称δ函数是在实数上定义的一个广义函数或分布。它在除零以外的点上都等于零，且其在整个定义域上的积分等于1。函数有时可看作是在原点处无限高、无限细，但是总面积为1的一个尖峰。

狄拉克δ函数是以零为中心的正态分布$\delta_a(x) = \dfrac{1}{a\sqrt{\pi}} e^{-x^2/a^2}$随$a \to 0$的（分布意义上的）极限。



eq17两个项的直观描述：第一项衡量**解码器**从变分分布中重构的概率，第二项衡量学到的变分分布与隐变量的先验分布的相似程度。因此，最大化ELBO等同于最大化其第一项并最小化其第二项。最小化第二项可以鼓励编码器学习一个优质的高斯分布，而不是崩溃成一个狄拉克 δ 函数。所以maximize ELBO作为代理任务既可以更好地学习编码器 $q_\theta(z \mid x)$，也可以更好地学习解码器 $p_\varphi(x \mid z)$。

VAE：如何在 δ 和 θ 上联合优化 ELBO

VAE的编码器通常选择具有对角协方差的多元高斯进行建模，而先验通常被选择为标准的多元高斯：
eq.18

$$q_\theta(z \mid x) \ = \ \mathcal{N}(z;\ \mu_\theta(x),\ \sigma_\theta^2(x) I)$$

eq.19

$$p(z) \ = \ \mathcal{N}(z;\ 0,\ I)$$

最后，ELBP的KL Divergence可以用解析计算，而重建项可以用蒙特卡洛估计来近似。
目标函数可以改写为(约等):
eq.20

$$arg\,max \sum_{l=1}^{L} log\, p_\phi(x \mid z^{(l)}) \ - \ D_{KL}(q_\theta(z \mid x) \,||\, p(z))$$

重参数化(fig5)

对于数据集中的数据x，其中隐变量 $\{\boldsymbol{z}^{(l)}\}_{l=1}^{L}$ 是从 $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ 中采样的。然而，在这个默认设置中出现了一个问题：计算损失的每个 $\boldsymbol{z}^{(l)}$ 都是随机抽样生成的，该过程通常是不可微的。幸运的是，当 $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ 被设计为对某些分布（例如多元高斯分布）进行建模时，可以通过重新参数化技巧（*reparameterization trick*）来解决这个问题。

重参数化技巧将随机变量重写为噪音变量的确定函数；这允许通过梯度下降优化非随机项。例如，从任意均值μ和方差 σ^2 的正态分布 $x \sim \mathcal{N}(x;\mu,\sigma^2)$ 采样得到的样本可以重写为：

$$x = \mu + \sigma\epsilon \quad \text{with } \epsilon \sim \mathcal{N}(\epsilon;0,\mathrm{I})$$