1、前言

大家好，今天我们讲Flow Matching，即流匹配。再讲解之前，您必须要了解之前讲过的两个概念——**神经常微分方程、连续归一化流**，因为Flow Matching就是建立在他们的基础上的

参考论文：[2210.02747] Flow Matching for Generative Modeling (arxiv.org)

参考代码：①Simple reimplementation of Flow Matching for Generative Modeling

②Repository search results

2、引入

在上期视频，我们讲了连续归一化流CNF，提到了其中的训练的方法与预测方法。从中我们不难发现，它里面存在一个比较严重的问题—— **在训练的时候，它需要模拟ODE数值求解运算，这将导致训练速度大幅下降**

为了解决这个问题，改论文提出了**Flow Matching**

在讲解Flow Matching之前，让我们回顾一下Neural ODE

2.1、Neural ODE

假设，存在一个微分方程

$$\frac{dx(t)}{dt} = f(x(t), t, \theta) \tag{1}$$

我们可以通过数值解法（比如欧拉法）

$$x(t+1) = x(t) + h \times f(x(t), t, \theta)$$

现在，为了采取与论文一致的符号，我们来对Eq.(1)进行符号上的修改

对Eq.(1)，将$x(t)$改写成$\phi_t(x)$，微分函数f改写成v，同时为了简便，我们省去参数θ，以及时刻t（或者说把它坍缩入$\phi_t(x)$当中），所以

$$\frac{d\phi_t(x)}{dt} = v(\phi_t(x)) \tag{2}$$

其中，$\phi_0(x) = x$，x为训练数据点$x = (x^1, x^2, \cdots, x^d) \in R^d$

现在，让我们对里面的量，使用更为数学化的表达，对于微分函数v，我们称为**向量场**

而$\phi_t(x)$可以视为t时刻对应的图像数据，即对**变量替换定理**，有

$$p_t = p_0(\phi_t^{-1}(x)) \det \left[\frac{\partial \phi_t^{-1}}{\partial x}(x) \right] \tag{3}$$

对满足Eq.(3)的向量场对应的ϕ，则称向量场v生成**概率密度路径**p_t（随时间变化而变化，所以在时间上是一条路径）

证明向量场v是否生成概率路径p_t的一种方法是使用连续性方程，即

$$\frac{d}{dt}p_t(x) + div(p_t(x)v_t(x)) = 0 \tag{4}$$

其中div是散度，即$div = \sum_{i=1}^{d} \frac{\partial}{\partial x^i}$

Eq.(4)怎么来的呢？如果你熟悉Fokker-Plank方程，想来一眼就可看出，这其实就是Fokker-Plank方程忽略掉扩散项得来的（不熟悉请参考什么是Fokker-Planck方程？，若无兴趣，直接记住该方程式即可，不影响下方推导）

Eq.(4)为充要条件，能够确保向量场v_t生成p_t，为了下方的使用，我们把散度提到等号右边

$$\frac{d}{dt}p_t(x) = -div(p_t(x)v_t(x)) \tag{4}$$

3、Flow Matching（FM）

3.1、优化目标

前面说到，如果使用CNF，那么我们每一次都要使用数值解法迭代去反向传播，这种训练速度显然是非常慢的，让我们来考虑一种更为简单的方法

设x_1表示某种未知数据分布$q(x_1)$分布的随机变量。我们假设我们只能获得来自$q(x_1)$的数据样本，但无法访问密度函数本身。

此外，我们让p_t是一个概率路径，p_0是一个简单的分布（比如，标准正态分布$p_0(x) = N(x|0, I)$，并让p_1 在分布上近似等于q。稍后我们将讨论如何构建这样的路径。然后设计流匹配目标来匹配此目标概率路径，这将允许我们从p_0流到p_1

$$L_{FM}(\theta) = \mathbb{E}_{t,p_t(x)}||v_t(x,\theta) - u_t(x)||^2 \tag{5}$$

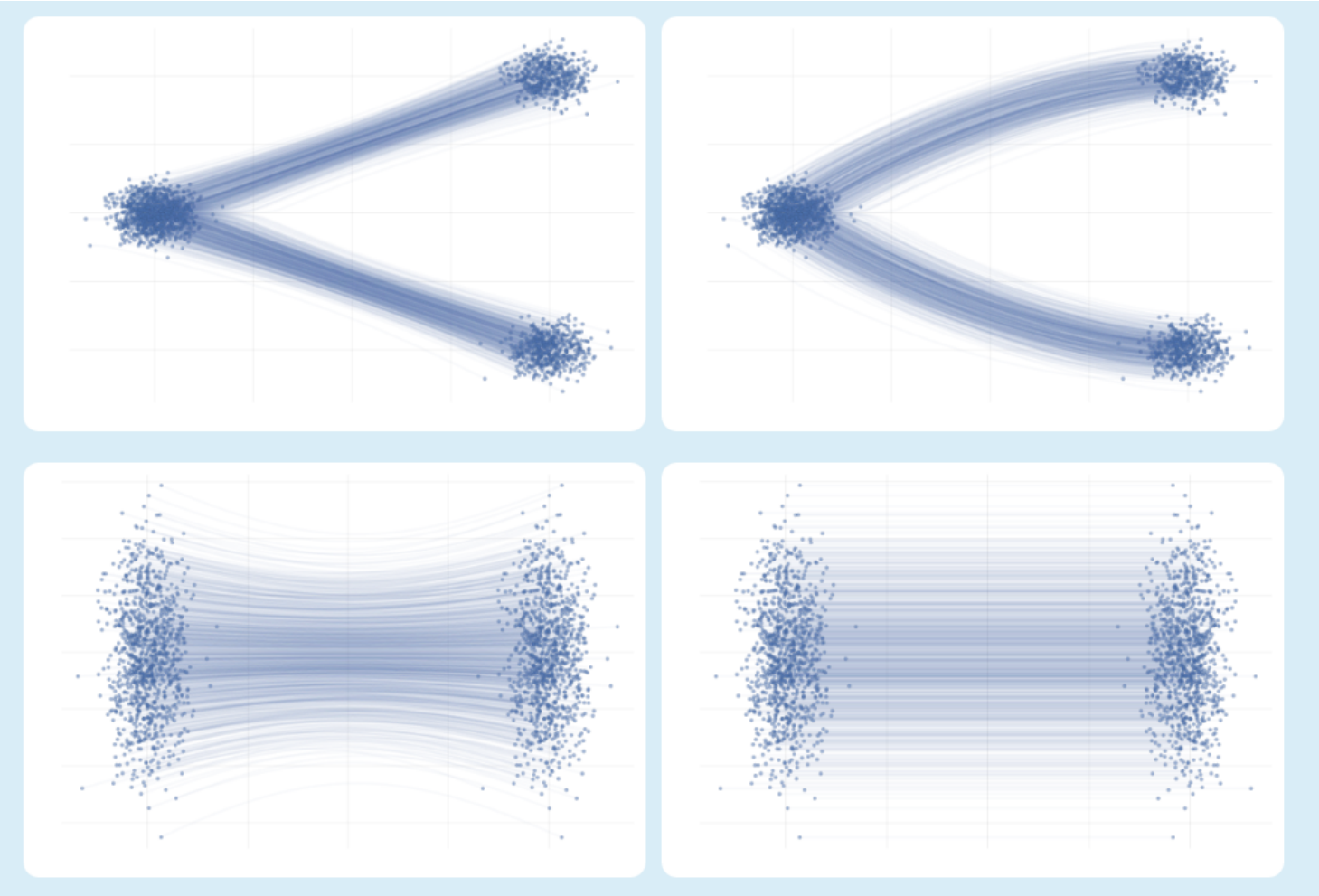
其中，向量场$u_t(x)$可生成对应的概率路径$p_t(x)$

也就是说，我们用一个神经网络去学习Eq.(4)中的$v_t(x,\theta)$，θ为神经网络的参数。p_t是一条概率路径，但一定要满足在p_0是简单分布，在p_1近似等于q

这样的话，如果我们训练完成，那么就可以利用学习到的向量场v，通过Eq.(2)，即可采样从一个简单分布中采样一个初值x_0，数值解法得到x_1，由于p_1近似等于q，所以就近似得到了数据样本

然而，这是建立在可以优化Eq.(5)的情况下的，现在我们的问题是，我们并不知道u_t、p_t是什么，根本没法训练

p_t的形式是不唯一的，我们有很多的路径满足这一要求



> 所以，现在的问题，就是想办法去构建u_t, p_t

3.2、从条件概率路径和向量场去构建p_t, u_t

假定一个条件概率路径$p_t(x|x_1)$，并满足$p_0(x|x_1) = p_0(x)$，而$p_1(x|x_1)$时，让$x_1 \approx x$，比如当$p_1(x|x_1) = N(x_1, \sigma^2 I)$，里面的$\sigma$足够小的情况下，$x_1 \approx x$

即让我们假定$p_t(x)$：

$$p_t(x) = \int p_t(x|x_1) q(x_1) dx_1 \tag{6}$$

对Eq.(6)，很显然，当t=1，则（上述曾定义过，t=1时，$p_1 \approx q$）

$$p_1(x) = \int p_1(x|x_1) q(x_1) dx_1 \approx q(x)$$

其中，$p_t(x|x_1)$所对应的向量场，我们记为$u_t(x|x_1)$，即

$$\frac{d}{dt} p_t(x|x_1) = -div\left(p_t(x|x_1) u_t(x|x_1)\right) \tag{7}$$

这样的话，我们可以利用连续性方程，反推出概率密度路径p_t所对应的向量场u_t的表达式

3.2.1、求证以Eq.(5)作为概率密度路径时的向量场u_t

假定被积函数满足莱布尼兹积分法则，则我们可交换微分和积分

由Eq.(3)的连续性方程，我们可以反推出Eq.(6)所对应的向量场

那么，则有

$$\begin{aligned}
\frac{d}{dt}p_t(x) &= \frac{d}{dt}\int p_t(x|x_1)q(x_1)dx_1 \\
&= \int \frac{d}{dt}\left(p_t(x|x_1)\right)q(x_1)dx_1 \\
&= \int -div\left(p_t(x|x_1)u_t(x|x_1)\right)q(x_1)dx_1 \\
&= -\int div\left(p_t(x|x_1)u_t(x|x_1)q(x_1)\right)dx_1 \\
&= -div\int p_t(x|x_1)u_t(x|x_1)q(x_1)dx_1 \\
&= -div\int \frac{u_t(x|x_1)p_t(x|x_1)q(x_1)}{p_t(x)}p_t(x)dx_1 \\
&= -div\left(p_t(x)\int \frac{u_t(x|x_1)p_t(x|x_1)q(x_1)}{p_t(x)}dx_1\right) \\
&= -div\left(p_t(x)u_t(x)\right)
\end{aligned}$$

从最后两个等号不难看出，有

$$u_t(x) = \int \frac{u_t(x|x_1)p_t(x|x_1)q(x_1)}{p_t(x)}dx_1 \tag{8}$$

我们去构建了Eq.(6)这样的概率路径，以及得到与之对应的Eq.(8)向量场，有什么用呢？

> 其实，如果我们这样去表示的话，整个计算会相对容易很多，何以见得？我们把未知的概率路径p_t和向量场u_t分解成了条件概率路径和条件向量场的表达式。相对于边缘概率路径p_t和向量场u_t，条件的形式显然更容易去计算，因为他们依赖于数据点x_1，而数据点x_1我们则是知道的

3.3、条件流匹配（CFM）

理论上，如果我们可以算出Eq.(6)和Eq.(8)，那么就可以进行优化更新了，然而，我们做不到！为何？

因为里面有个积分，积分我们是处理不了的，因此，我们考虑一个条件流匹配（CFM）

$$L_{CFM}(\theta) = \mathbb{E}_{t,q(x_1),p_t(x|x_1)}\left[||v_t(x,\theta) - u_t(x|x_1)||^2\right] \tag{9}$$

我们可以证明L_{CFM}的优化目标与L_{FM}是一致的，而Eq.(9)显然是相对容易优化的，因为它里面只有条件概率和条件向量场，而无需求边缘概率和向量场，也就无需再进行积分运算了

3.3.1、证明L_{CFM}与L_{FM}目标一致性

先来看L_{FM}的训练目标

$$
\begin{aligned}
L_{FM}(\theta) =& \mathbb{E}_{t,p_t(x)}||v_t(x,\theta)-u_t(x)||^2 \\
=& \mathbb{E}_{t,p_t(x)}\left[||v_t(x,\theta)||^2-2v_t(x,\theta)^Tu_t(x)+||u_t(x)||^2\right] \\
=& \underbrace{\mathbb{E}_{t,p_t(x)}\left[||v_t(x,\theta||^2\right]}_{①}-\underbrace{\mathbb{E}_{t,p_t(x)}\left[2v_t(x,\theta)^Tu_t(x)\right]}_{②}+\underbrace{\mathbb{E}_{t,p_t(x)}\left[||u_t(x)||^2\right]}_{\text{与}\theta\text{无关}}
\end{aligned}
$$

接着看$L_{CFM}(\theta)$的

$$
\begin{aligned}
L_{CFM}(\theta) =& \mathbb{E}_{t,q(x_1),p_t(x|x_1)}\left[||v_t(x,\theta)-u_t(x|x_1)||^2\right] \\
=& \underbrace{\mathbb{E}_{t,q(x_1),p_t(x|x_1)}\left[||v_t(x,\theta)||^2\right]}_{①}-\underbrace{\mathbb{E}_{t,q(x_1),p_t(x|x_1)}\left[2v_t(x,\theta)^Tu_t(x|x_1)\right]}_{②}+\underbrace{\mathbb{E}_{t,q(x_1),p_t(x|x_1)}\left[||u_t(x|x_1)||^2\right]}_{\text{与}\theta\text{无关}}
\end{aligned}
$$

这样的话，其实我们只需要证明前面两项相等即可（先看L_{FM}的第一项，为书写简便，我们暂时忽略掉关于t的积分）

$$
\begin{aligned}
\mathbb{E}_{t,p_t(x)}\left[||v_t(x,\theta||^2\right] =& \int||v_t(x,\theta)||^2p_t(x)dx \\
=& \int||v_t(x,\theta)||^2p_t(x|x_1)q(x_1)dx_1dx \\
=& \mathbb{E}_{q(x_1),p_t(x|x_1)}\left[||v(x,\theta)||^2\right]
\end{aligned}
$$

容易看到如果把时间t加入回去，那么刚好等于L_{CFM}里面的第一项

再来看L_{FM}的第二项（依然暂时忽略掉t的积分）

$$
\begin{aligned}
\mathbb{E}_{t,p_t(x)}\left[2v_t(x,\theta)^Tu_t(x)\right] =& \int 2v_t(x,\theta)^Tu_t(x)p_t(x)dx \\
=& \int 2v_t(x,\theta)^T\int\frac{u_t(x|x_1)p_t(x|x_1)q(x_1)}{p_t(x)}dx_1p_t(x)dx \\
=& \int 2v_t(x,\theta)^T\int\frac{u_t(x|x_1)p_t(x|x_1)q(x_1)}{p_t(x)}p_t(x)dx_1dx \\
=& \int 2v_t(x,\theta)^T\int u_t(x|x_1)p_t(x|x_1)q(x_1)dx_1dx \\
=& \iint 2v_t(x,\theta)^Tu_t(x|x_1)p_t(x|x_1)q(x_1)dx_1dx \\
=& \mathbb{E}_{q(x_1),p(x|x_1)}\left[2v_t(x,\theta)^Tu_t(x|x_1)\right]
\end{aligned}
$$

把时间t加入回去，那么结果刚好等于L_{CFM}的第二项

由此，得证L_{CFM}与L_{FM}目标一致

3.4、条件概率路径和向量场

考虑到Eq.(9)的优化目标，为了真正应用起来，我们仍然需要为条件概率路径和条件向量场选择具体的形式

比如，我们假设条件概率路径是高斯分布，即

$$p_t(x|x_1) = N(x|\mu_t(x_1), \sigma_t(x_1)^2 I) \tag{10}$$

其中，$\mu : [0,1] \times \mathbb{R}^d \to \mathbb{R}^d$是高斯分布的均值，随着时间变化而变化，$\sigma$同理。根据前面的的假设需求，有$\mu_0(x_1) = 0, \sigma_0(x_1) = 1$，也就是在t=0时刻，$p_0(x|x_1) = N(x|0, I)$；而在时刻t=1，则$\mu_1(x_1) = x_1, \sigma_1(x_1) = \sigma_{\min}$，如之前所说那般，$\sigma_{\min}$是一个极其小的量

让我们考虑把把高斯分布作为一个流变换（以x_1为条件），即

$$\psi_t(x) = \sigma_t(x_1)x + \mu_t(x_1) \tag{11}$$

如果把x当作是标准高斯分布（对应x_0），那么整个变化就会变成高斯分布的重参数化形式，也就是相当于从Eq.(10)采样出数据点

有了上述表达，我们便可以得到向量场的微分表达

$$\frac{d}{dt}\psi_t(x) = u_t(\psi_t(x)|x_1)$$

通过x_0重参数化$p_t(x|x_1)$，可得

$$\begin{aligned} L_{CFM}(\theta) = \mathbb{E}_{t,q(x_1),p_t(x|x_1)}\left[||v_t(x,\theta) - u_t(x|x_1)||^2\right] =& \mathbb{E}_{t,q(x_1),p_0(x_0)}\left[||v_t(\psi(x_0),\theta) - u_t(\psi(x_0)|x_1)||^2\right] \\ =& \mathbb{E}_{t,q(x_1),p_0(x_0)}\left[||v_t(\psi(x_0),\theta) - \frac{d}{dt}\psi_t(x_0)||^2\right] \end{aligned}$$

如果一切都按照我们定义的那般，那么我们可以得出对应的向量场为

$$u_t(x|x_1) = \frac{\sigma_t'(x_1)}{\sigma_t(x_1)}(x - \mu_t(x_1)) + \mu_t'(x_1) \tag{12}$$

3.4.1、证明Eq.(12)

为了简单性，让我们考虑用$w_t(x) = u_t(x|x_1)$

对于Eq.(11)，显然是可逆的，我们记

$$\psi_t(x) = \sigma_t(x_1)x + \mu_t(x_1) = y \rightarrow x = \frac{y - \mu_t(x_1)}{\sigma_t(x_1)} = \psi_t^{-1}(y) \tag{13}$$

所以

$$\frac{d}{dt}\psi_t(x) = \frac{d}{dt}\psi_t(\psi_t^{-1}(y)) = w_t(y)$$

对Eq.(11)关于t求导

$$\frac{d}{dt}\psi_t(x) = \sigma_t'(x_1)x + \mu_t'(x_1) = w_t(x)$$

由Eq.(13)，将上式的x代换掉，则

$$w_t(y) = \sigma_t'(x_1)\frac{y - \mu_t(x_1)}{\sigma_t(x_1)} + \mu_t'(x_1) = \frac{\sigma_t'(x_1)}{\sigma_t(x_1)}(y - \mu_t(x_1)) + \mu_t'(x_1)$$

把y替换回x后，即可得到

$$w(x) = u_t(x|x_1) = \frac{\sigma_t'(x_1)}{\sigma_t(x_1)}(x - \mu_t(x_1)) + \mu_t'(x_1) \tag{12}$$

得证Eq.(12)

4、实例应用

4.1、高斯条件概率路径的特殊实例（VE、VP）

让我们考虑（实际上就是我们以前讲过的VE SDE）

$$p_t(x|x_1) = N(x|x_1, \sigma_{1-t}^2 I)$$

其中σ_t 是递增函数，$\sigma_0 = 0$和$\sigma_1 \gg 1$。$\mu_t(x_1) = x_1$，$\sigma_t(x_1) = \sigma_{1-t}$

根据Eq.(12)，我们即可得出对应的条件向量场

$$u_t(x|x_1) = -\frac{\sigma_{1-t}'(x_1)}{\sigma_{1-t}(x_1)}(x - x_1)$$

让我们考虑（实际上就是我们以前讲过的VP SDE）

$$p_t(x|x_1) = N\left(x|\alpha_{1-t}x_1, (1 - \alpha_{1-t}^2)I\right)$$

其中，$\alpha_t = e^{-\frac{1}{2}T(t)}$，$T(t) = \int_0^1 \beta(s)ds$，$\beta$是噪声尺度。$\mu_t(x_1) = \alpha_{1-t}x_1, \sigma_t(x_1) = \sqrt{1-\alpha_{1-t}^2}$

根据Eq.(12)，我们即可得出对应的向量场

$$u_t(x|x_1) = \frac{\alpha'_{1-t}}{1-\alpha_{1-t}^2}(\alpha_{1-t}x - x_1) = -\frac{T'(1-t)}{2}\left[\frac{e^{-T(1-t)}x - e^{-\frac{1}{2}T(1-t)}x_1}{1-e^{-T(1-t)}}\right]$$

论文提到，由于这些概率路径之前是作为扩散过程的解推导出来的，因此它们实际上不会在有限时间内达到真正的噪声分布。在实践中，$p_0(x)$只是通过适当的高斯分布进行近似，用于采样和似然估计

论文提出了一种更稳定、更具鲁棒性的方法

4.2、最优传输条件向量场

让我们考虑

$$p_t(x|x_1) = N\left(x|tx_1, (1-(1-\sigma_{\min})t)^2 I\right)$$

即$\mu_t(x) = tx_1, \sigma_t(x) = 1-(1-\sigma_{\min})t$

既如此，我们也可得到对应的流变换（重参数化）

$$\psi_t(x) = (1-(1-\sigma_{\min})t)x + tx_1 \tag{14}$$

依据Eq.(14)，可以导出对应的条件向量场

$$u_t(x|x_1) = \frac{x_1 - (1-\sigma_{\min})x}{1-(1-\sigma_{\min})t}$$

并且，依据Eq.(14)，有

$$\frac{d}{dt}\psi_t(x_0) = x_1 - (1-\sigma_{\min})x_0$$

那么优化目标就可以化简成

$$\begin{aligned} L_{CFM}(\theta) =& \mathbb{E}_{t,q(x_1),p_0(x_0)}\left[||v_t(\psi(x_0),\theta) - \frac{d}{dt}\psi_t(x_0)||^2\right] \\ =& \mathbb{E}_{t,q(x_1),p_0(x_0)}\left[||v_t(\psi(x_0),\theta) - (x_1 - (1-\sigma_{\min})x_0)\,||^2\right] \end{aligned}$$

这就是我们该条件下的优化目标

5、结束

好了，本篇文章到此为止，如有问题，还望指出，阿里嘎多！

6、参考

流匹配简介 ·剑桥 MLG 博客 (cam.ac.uk)

Flow Matching For Generative Modeling-CSDN博客

Flow matching文献阅读（一） - 知乎 (zhihu.com)