理解扩散模型：一个统一的视角

ar

Xi

v:

22

08

.1

19

70

v1

[c

s.

LG

]2

**罗嘉文**

谷歌研究，大脑团队

calvin luo@google.com

2022年8月26日

**内容**

**简介：生成模型**  [**1**](#bookmark1)

**背景：** **ELBO、VAE和层次化的VAE**  [**2**](#bookmark2)

**证据下界** [**2**](#bookmark3)

**变分自动编码器**  [**4**](#bookmark4)

**分层变分自动编码器** [**5**](#bookmark5)

**变分扩散模型** [**6**](#bookmark6)

**学习扩散噪声参数**  [**14**](#bookmark7)

**三等效解释**  [**1** **5**](#bookmark8)

**基于分数的生成模型** [**17**](#bookmark9)

**指导**  [**20**](#bookmark10)

**分类器指导**  [**21**](#bookmark11)

**无分类器指导** [**21**](#bookmark12)

**关闭**  [**22**](#bookmark13)

**简介：生成模型**

给定从一个感兴趣的分布中观察到的样本x，生成模型的目标是学习建模其真实的数据分布p (x)。一旦学习到， 我们就可以随意地从我们的近似模型中生成新的样本。此外，在一些公式下，我们也可以使用学习模型来评估观 测或抽样数据的可能性。

在目前的文献中有几个众所周知的方向，我们只会在高水平上简要介绍。生成对抗性网络（GANs）对复杂分布的 抽样过程进行建模，并以对抗性的方式进行学习。另一类生成模型，称为“基于可能性的 ”，试图学习一个为观 察到的数据样本分配高可能性的模型。这包括自回归模型、规范化流和变分自动编码器（VAEs）。另一种类似的 方法是基于能量的建模，其中一个分布被学习为一个任意灵活的能量函数，然后被归一化。

基于分数的生成模型高度相关；他们不是学习建模能量函数本身，而是学习基于能量的模型的分数作为一个神经 网络。在这项工作中，我们探索和回顾了扩散模型，正如我们将证明的那样，它有基于可能性和基于分数的解释 。我们以惊人的细节展示了这些模型背后的数学原理， 目的是让任何人都可以跟踪和理解什么是扩散模型以及它 们是如何工作的。

**背景：** **ELBO、VAE和层次化的VAE**

对于许多模式，我们可以认为我们观察到的数据是由一个相关的看不见的潜在变量表示或生成的，我们可以用随 机变量z来表示它。表达这一观点的最好的直觉是通过柏拉图的洞穴寓言。在寓言中，一群人一生都被锁在一个 洞穴里，只能看到投射在他们前面的一堵墙上的二维阴影，这些阴影是由看不见的三维物体经过火灾前产生的。 对这些人来说，他们所观察到的一切实际上都是由他们永远无法看到的高维抽象概念所决定的。

类似地，我们在现实世界中遇到的对象也可以作为一些高级表示的函数生成；例如，这种表示可能封装抽象属性 , 如颜色、大小、形状等。然后，我们观察到的可以被解释为这些抽象概念的三维投影或实例化，就像洞穴人们 观察到的实际上是三维物体的二维投影一样。虽然洞穴里的人永远看不到（甚至完全理解）隐藏的物体，但他们 仍然可以对它们进行推理和推断； 以类似的方式，我们可以近似地描述我们观察到的数据的潜在表征。

虽然柏拉图的寓言说明了潜在变量背后的观点，即决定观察，但这个类比需要注意的是，在生成建模中，我们通 常寻求学习低维的潜在表征，而不是高维的表征。这是因为试图学习比观察更高维度的表示，没有强先验是徒劳 的努力。另一方面，学习低维延迟也可以被视为一种压缩的形式，并可以潜在地揭示描述观察的语义上有意义的 结构。

**证据下界**

在数学上，我们可以想象潜在的变量和我们观察到的数据是由一个联合分布p（x，z）建模的。回想一下，生成 建模的一种方法，称为“基于可能性的 ”，是学习一个模型，以最大化所有观察到的x的可能性p (x)。我们有两

种方法可以操纵这种联合分布来恢复纯粹我们的观测数据p (x)的可能性；我们可以明确地边缘化潜在的

变量z：

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

(1)

或者，我们也可以求助于概率的链式规则：

p (x) =

p (x, **z)**

p (z|x)

(2)

直接计算和最大化似然p (x)是困难的，因为它要么涉及到积分方程1中的所有潜在变量z，这对于复杂模型是棘 手的，要么涉及到访问方程2中的地面真实潜在编码器p（z|x） 。然而，利用这两个方程，我们可以推导出一个 称为证据下界（ELBO）的术语，正如其名称所示，它是证据的下界。在这种情况下，证据被量化为观测数据的对 数似然值。然后，最大化ELBO成为优化潜在变量模型的代理目标；在最好的情况下，当ELBO被强大地参数化和完 美地优化时，它就变得完全等价于证据。在形式上，ELBO的方程为：

[ ]

Eq (z|x) 记录

(3)

为了明确与证据的关系，我们可以从数学上写下：

[ ]

日志p (x) Eq (z|x)· 记录 (4)

这里，q（z|x）是一个灵活的近似变分分布，我们寻求优化。直观地说，它可以被认为是一个可参数的模型， 它可以用来估计给定观测x的潜在变量的真实分布；换句话说，它寻求近似真实的后验p（z|x）。正如我们在探 索变分自动编码器时将看到的，当我们通过调整参数来最大化ELBO来增加下界时，我们可以访问可用于建模真实

数据分布并从中采样的组件，从而学习生成模型。现在，让我们试着深入研究一下

来了解为什么ELBO是一个我们想要最大化的目标。

让我们从用方程1推导ELBO开始：

\

（应用公式1） (5)

日志p (x) = 日志 p (x, z)dz

\

= 日志 dz （乘以1 = ) (6)

[ ]

= 日志Eq (z|x)·  （期望的定义） (7)

[ ]

Eq (z|x) 记录 （应用延森不等式） (8)

在这个推导中，我们通过应用延森不等式直接得到了我们的下界。然而，这并没有为我们提供太多关于罩下实际 发生的有用信息；至关重要的是，这个证据并没有直观地解释为什么ELBO实际上是证据的下界，因为Jensen的不 等式处理了它。此外，仅仅是知道ELBO确实是数据的一个下界，并不能真正告诉我们为什么我们要将它作为一个 目标来最大化。为了更好地理解证据和ELBO之间的关系，让我们进行另一个推导，这次使用方程2：

\

日志p (x) = 日志p (x) q (z|x)dX

=\q (z|x)(log p

( ))

= Eq (z|x) [日志 p (x)]

[

]

= Eq (z|x) 记录

[

]

]

= Eq (z|x) 记录

[

[

= Eq (z|x) 记录

+ Eq (z|x) 记录

] [ ]

= E q (z|x) 记录 + D氯化钾

[

 (q (z|x) | p (z|x)) Eq (z|x) 记录

]



\

(乘以1= q (z|x)dz)

（将证据纳入积分）（期望的

定义）（应用公式2）

（乘以1 = )

（分割期望）（KL散度的定义

)

(KL差异始终为0）

(9)

(10) (11)

(12)

(13) (14) (15)

(16)

从这个推导中，我们从公式15清楚地观察到，证据等于近似的后q（z|x）和真正的后p（z|x）之间的ELBO加上KL 散度。事实上，正是这个KL散度项被第一次推导的等式8中的延森不等式神奇地消除了。理解这一术语不仅是理 解ELBO和证据之间关系的关键，也是理解优化ELBO是一个合适的目标的原因。

首先，我们现在知道为什么ELBO确实是一个下界：证据和ELBO之间的差异是一个严格的非负的KL项，因此ELBO的 值永远不能超过证据。

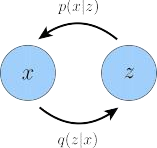


图1：一个以图形方式表示的变分自动编码器。这里，编码器q（z|x）定义了观测值x的潜在变量z上的分布，而p （x|z）将潜在变量解码为观测值。

其次，我们探讨为什么我们寻求最大化ELBO。在引入了我们想要建模的潜在变量z之后，我们的目标是学习描述 我们观察数据的潜在结构。换句话说，我们想要优化我们的变分后验q（z|x）的参数，以精确匹配真实的后验 分布p（z|x），这是通过最小化它们的KL散度（理想情况下为零）来实现的。不幸的是，直接最小化这个KL发散 项是很困难的，因为我们无法获得地面真相p（z|x）分布。然而，注意左边的方程15，我们的数据的可能性( 因此我们的证据项日志p (x)）总是一个常数，因为它是计算边缘化所有延迟z联合分布p（x，z），不依赖于任 何。由于ELBO和KL散度项之和为一个常数，因此关于ELBO项的任何最大化都必然会调用KL散度项的相等最小化 。因此，ELBO可以被最大化，作为学习如何完美地建模真实的潜在后验分布的代理；我们优化的ELBO越多，我 们的近似后验就越接近真实的后验。此外，一旦训练，ELBO也可以用来估计观察到的或生成的数据的可能性，因 为它是学习近似模型证据日志p (x)。

**变分自动编码器**

在变分自动编码器（VAE） [1]的默认公式中，我们直接最大化ELBO。这种方法是变分的，因为我们优化了一个 参数化的潜在后验分布家族中的最佳q（z|x）。它被称为自动编码器，因为它让人想起传统的自动编码器模型 , 在该模型中，输入数据在经过中间瓶颈表示步骤后被训练来预测自己。为了明确这种联系，让我们进一步分析 ELBO术语：

] [

[

Eq (z|x) 记录

]

= Eq (z|x) 记录

[ ]

= Eq (z|x) [日志 p**e**(x|z)] + Eq (z|x) 记录

= Eq (z|x) [日志 p**e**(x|z)] - D氯化钾 (q (z|x) | p (z))

" –企 物 " –企 物

重建项前一个匹配项

（概率链规则）（17）

（分割期望值）（18）（KL散度的

定义）（19）

在这种情况下，我们学习了一个中间瓶颈分布q（z|x），它可以被视为一个编码器；它将输入转换为可能延迟的 分布。同时，我们学习了一个确定性函数p**e**（x|z）将给定的潜在向量z转换为观测x，可以解释为解码器。

方程19中的两项都有直观的描述：第一项测量从我们的变分分布中解码器的重构可能性；这确保了学习到的分布 是建模原始数据可以再生的有效延迟。第二项测量了学习到的变分分布与对潜在变量所持有的先验信念的相似程 度。最小化这个术语会鼓励编码器实际上学习一个分布，而不是坍缩成一个狄拉克函数。因此，最大化ELBO相当 于最大化它的第一项和最小化其第二项。

VAE的一个定义特性是如何在参数和参数上联合优化ELBO。θ编码器

中的VAE通常被选择来建模一个具有对角协方差的多元高斯，并且先验通常被选择为一个标准的多元高斯：

q (z|北) = N(z ; u (北), (北)I) ([20)](#bookmark14)

p (z) = N(z ; 0, I) [(21)](#bookmark15)

然后，可以解析地计算ELBO的KL散度项，重构项可以用蒙特卡罗估计来近似。然后，我们的目标可以重写为：

L

arg max Eq (z|x) [日志 p（北|z）] - D 氯化钾  (q (z|北) | p (z))arg最大值工日志p(北 |z(l)) - D氯化 钾 (q (z|北) | p (z)) (22)

, l=1 ,

潜伏期(l)}从q（z|北）中采样，对于数据集中的每个观测北。然而，在这个默认设置中出现了一个问题： 每个z( l) 我们的损失计算是由一个随机抽样过程产生的，它通常是不可微的。幸运的是，当q（z|北）被设计用 来建模某些分布，包括多元高斯分布时，这可以通过重新参数化技巧来解决。

重参数化技巧将一个随机变量重写为一个噪声变量的确定性函数；这允许通过梯度下降来优化非随机项。例如， 来自正态分布x N (x；u，a2)，具有任意的均值u和方差a2可重写为：

=u+（=；0，+）~

换句话说，任意高斯分布可以解释为标准高斯分布（其中e是一个样本），它们的均值从0移到目标均值u，它们 的方差被目标方差a拉伸2 .因此，通过重新参数化技巧，可以从任意高斯分布中采样，从一个标准高斯分布中采 样，根据目标标准差进行缩放，并根据目标均值移动。

因此，在VAE中，每个z被计算为输入北和辅助噪声变量e：z=u（北）+（北）勺e与e N（e；0,0，勺）的

确定性函数~

其中，勺表示一个元素级的产品。在这个重新参数化的z版本下，然后可以根据需要计算梯度，以优化u和。 因此，VAE利用重新参数化技巧和蒙特卡罗估计来联合优化ELBO在和。θ

在训练一个VAE后，可以通过直接从潜在空间p (z)中采样，然后通过解码器运行它来生成新的数据。当z的维数 小于输入北的维数时，变分自动编码器特别有趣，因为我们可能正在学习紧凑的、有用的表示。此外，当学习到 语义上有意义的潜在空间时，可以在将潜在向量传递给解码器之前进行编辑， 以更精确地控制生成的数据。

**分层变分自动编码器**

层次变分自动编码器（HVAE） [2,3]是一种VAE的推广，它扩展到潜在变量上的多个层次。在这个公式下，潜在变 量本身被解释为来自其他更高层次、更抽象的潜在变量。直观地说，就像我们把我们的三维观察对象视为由更高 层次的抽象潜在产生的一样，柏拉图洞穴中的人把三维对象视为产生二维观察的潜伏。因此，从柏拉图的洞穴居 民的角度来看，他们的观察可以被视为由深度2（或更多）的潜在层次来建模。

然而，在具有T层次的一般HVAE中，每个潜在的都允许限制所有以前的潜伏期，在本工作中，我们关注一种特殊 情况，我们称之为马尔可夫HVAE（MHVAE）。在MHVAE中，生成过程是一个马尔可夫链；也就是说，每个层次结构 的过渡都是马尔可夫的

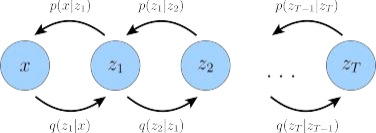


图2：一个具有T个层次延迟的马尔可夫层次变分自动编码器。生成过程被建模为一个马尔可夫链，其中每个潜在 的zt仅由之前的潜在z生成t+1.

解码每个潜在zt只有对先前潜在z的条件t+ 1 .在直观上和视觉上，这可以看作是简单地相互叠加，如图2所示； 另一个适当的术语描述

这个模型是一个递归的VAE。在数学上，我们将一个马尔可夫HVAE的联合分布和后验表示为：

T

p (x, z1:T) = p (z T)p**e**(x|z 1) u p **e**(z t — 1|z t)

t=2

T

q (z1:T|x) = q (z 1|x) u q (z t|z t — 1)

t=2

然后，我们可以很容易地扩展ELBO为：

|  |  |
| --- | --- |
| = 日志\p (x、  日志p (x)  =  \  日志  dz1 : T = 日志 [ ]  Eq (z 1:T|x)  [ ]  Eq (z 1:T|x) 记录 | （应用公式1）  （乘以1 = )  （期望的定义）  （应用延森不等式） |

然后，我们可以将我们的联合分布（式23)和后式（式24)插入式28，以生成一种替代形式：

Eq (z·1:T|x) 记录 = Eq (z 1:T|x) 记录

[ ] [ ]

正如我们下面将展示的，当我们研究变分扩散模型时，这个目标可以进一步分解为可解释的组件。

(23)

(24)

(25)

(26)

(27)

(28)

(29)

**变分扩散模型**

考虑变分扩散模型（VDM） [4,5,6]的最简单的方法是作为一个马尔可夫层次的变分自动编码器，有三个关键的限 制：

潜在维度完全等于数据维度

不学习潜在编码器在每个时间步长的结构；它被预先定义为一个线性高斯模型。换句话说，它是一个以前一 个时间步长的输出为中心的高斯分布

潜在编码器的高斯参数随时间变化，使得潜在编码在最终时间步长T的分布是标准高斯分布

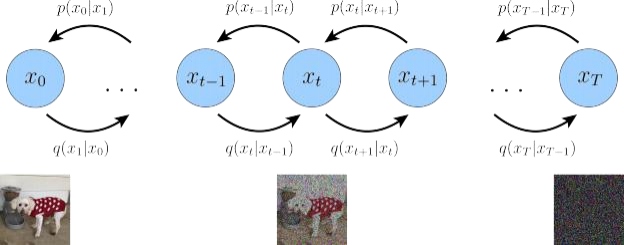


图3：一个变分扩散模型的可视化表示0表示真实的数据观察，如自然图像，xT表示纯高斯噪声，和xt是x的中间 噪声版本吗0.每个q (xt|xt — 1)被建模为一个高斯分布，它使用前一个状态的输出作为其平均值。

此外，我们明确地保持了从一个标准的马尔可夫层次变分自编码器的层次转换之间的马尔可夫性质。

让我们扩展一下这些假设的含义。从第一个限制开始，通过滥用一些符号，我们现在可以将真实数据样本和潜在 变量表示为xt，其中t = 0表示真实数据样本，t 2 [1，T]表示对应的潜在数据，其层次索引为t。VDM后验与

MHVAE后验相同（式24），但现在可以改写为：

T

q (x1:T|x0) =u t=1

q (xt|xt — 1)

(30)

从第二个假设中，我们知道编码器中每个潜在变量的分布都是一个围绕其先前的层次潜在变量为中心的高斯分布 。与马尔可夫HVAE不同，编码器在每个时间步长t的结构不被学习；它是固定为一个线性高斯模型，其中均值和 标准

偏差可以预先设置为超参数[5]，也可以学习为参数[6]。我们参数化

均值为u的高斯编码器t(xt) = ^ α t**x**t — 1，方差对t(xt) = (1 α-t我，在那里的形式

系数的选择使潜在变量的方差保持在相似的尺度上；换句话说，

编码过程是保持方差的。请注意，允许进行交替的高斯参数化，

并导致类似的推导。最主要的结论是， αt有一个（可能是可学习的）系数可以吗

为了灵活性，随层次深度t而变化。在数学上，编码器的转换被表示为：

 q (xt|xt — 1) = N(xt ; ^ αt**x**t — 1, (1 α-t)I) (31)

从第三个假设中，我们知道了αt随着时间的推移，根据一个固定的或可学习的时间表结构，这样最终潜在p (x 的分布T)是一个标准的高斯分布。然后，我们可以更新一个马尔可夫HVAE的联合分布（方程23），将一个VDM的 联合分布写为：

T

p (x0:T) = p (xT) u

t=1

在哪里

p (xT) = N(xT; 0,

p**e**(xt — 1|xt)

I)

(32)

(33)

总的来说，这组假设所描述的是随着时间的推移，图像输入的稳定噪声；我们通过添加高斯噪声来逐步破坏图像 , 直到最终它变得与纯高斯噪声完全相同。从视觉上看，这个过程如图3所示。

请注意，我们的编码器分布q (xt|xt — 1)不再被参数化，因为它们完全被建模为在每个时间步长上具有定义 的均值和方差参数的高斯分布。因此，在VDM中，我们只对学习条件p感兴趣**e**(xt — 1|xt)，这样我们就可以 模拟新的数据。在优化VDM后，采样过程与从p (x中采样高斯噪声一样简单T)和迭代地运行去噪转换p**e**(xt — 1|xt)为T步骤生成一个新的x0.

像任何HVAE一样，VDM可以通过最大化ELBO来优化，这可以推导为：

日志p (x) = 日志\p (x0:T)dx1:T(34)

\

= 日志 dx1:T(35)

[ ]

= 日志Eq (x1:T|x0)  (36)

[ ]

Eq (x1:T|x0) 记录 (37)

[ ]

= Eq (x1:T|x0) 记录 (38)

[ ]

= Eq (x1:T|x0) 记录 (39)

[ ]

= Eq (x1:T|x0) 记录 (40)

[ ] [ ]

= Eq (x1:T|x0) 记录 + Eq (x1:T|x0) 记录 (41)

= Eq (x1:T|x0)[日志p**e**(x0|x 1)] + Eq (x1:T|x0) 记录 + Eq (x1:T|x0) 记录 (42)

[ ] [ ]

[ ] [

= Eq (x1:T|x0)[日志p**e**(x0|x1)] + Eq (x1:T|x0) 记录 + Eq (x1:T|x0) 记录

]

 (43)

[ ] [ ]

= E q (x1|x0)[日志p**e**(x0|x 1)] + Eq (xT — 1,xT|x0) 记录 + E q (xt — 1,xt,xt+1|x0) 记录 (44)

= Eq (x1|x0)[日志pe(x0|x1)] - Eq (xT — 1|x0)[D氯化钾(q (xT|xT — 1) | p (xT))]

(45)

石

别

重建项前一石个配项

石企

别 石

之T — 1

- Eq (xt — 1 ,xt + 1 |x0 ) [D氯化钾 (q (xt |xt — 1 ) | pe (xt |xt+1))] t=1石 石企别

含糊的术语enc

ELBO的派生形式可以用其单个组件来解释：

1 . Eq(x1|x0)[日志p**e**(x0|x 1)]可以解释为一个重构项，预测给定第一步潜在性的原始数据样本的对数概率 。这个术语也出现在一个普通的VAE中，并且可以进行类似的训练。

2.Eq (xT — 1|x0)[D氯化钾(q (xT|xT — 1) | p (xT是一个先验匹配项；当最终的潜在分布与高斯先验相 匹配时，它被最小化。这个项不需要优化，因为它没有可训练的参数；此外， 由于我们假设一个足够大的T , 最终分布是高斯分布，这个项有效地变成零。

3.Eq(xt — 1,xt+1|x0)[D氯化钾(q (xt|xt — 1) | p**e**(xt|xt+1是一个一致性术语；它努力使x处的分布 t一致的，从向前和后退的过程。也就是说，对于每个中间时间步长，噪声图像的去噪步长应该与干净图像

的相应噪声步长相匹配；这在数学上反映在KL发散上。当我们训练p时，这个项被最小化了e(xt|xt+1)来 匹配高斯分布q (xt|xt — 1)，其定义见等式31。

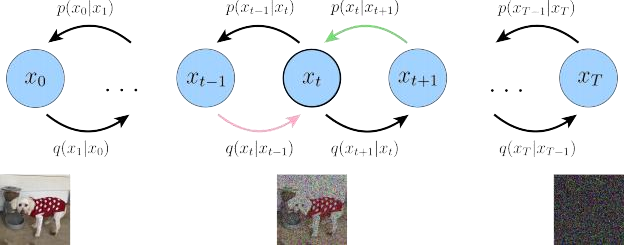


图4：在我们的第一次推导中，VDM可以通过确保每个中间x来优化VDMt，从它上面的潜伏期开始，p**e**( xt| xt + 1) 匹配潜在的它之前的高斯腐败t|xt — 1).在这个图中，对于每个中间的xt，我们最小化由粉色和绿色箭头表示 的分布之间的差异。

从视觉上看，对ELBO的这种解释如图4所示。优化一个VDM的成本主要由第三项控制，因为我们必须优化所有的时 间步长t。

在这个推导下，ELBO的所有项都被计算为期望，因此可以使用蒙特卡罗估计来近似。然而，实际上使用我们刚刚 推导出的术语来优化ELBO可能是次优的；因为一致性项是作为对两个随机变量{x的期望来计算的t — 1, xt+1} 对于每个时间步长，其蒙特卡罗估计的方差可能会高于每个时间步长仅使用一个随机变量估计的一项。 由于它是 通过对T - 1一致性项的相加计算出来的，ELBO的最终估计值对于较大的T值可能有较大的方差。

相反，让我们尝试为我们的ELBO推导出一种形式，其中每个项都被计算为一次只对一个随机变量的期望。关键的 见解是，我们可以将编码器转换重写为q (xt|xt — 1) = q (xt|xt — 1, x0)，其中由于马尔可夫性质，额外 的条件反射项是多余的。然后，根据贝叶斯规则，我们可以将每个转换改写为：

q (xt|xt — 1, x0) =

q (xt —

有了这个新的方程，我们可以重试从方程37中的ELBO恢q t的—推1 t：, **x**0 )q (xt

[ ]

日志p (x) Eq (x1:T|x0) 记录

[ ]

= Eq (x1:T|x 0) 记录

[ ]

= Eq (x1:T|x 0) 记录

[ ]

= Eq (x1:T|x 0) 记录

[ ]

= Eq (x1:T|x0) 日志+日志

2 [3](#bookmark16)

= Eq (x1:T|x0)4 日志+日志5

(46)

(47)

(48)

(49)

(50)

(51)

(52)

3

2

4

5 (53)

~~p~~

t—

日志+日志

= Eq (x1:T|x0)

[

]

= Eq (x1:T|x0) 记录

~~1)~~

一~~x1~~

**~~xT~~**

~~pe~~

一~~x1~~



+日志

+日志

]

(54)

[

= Eq (x1:T|x0) 日志+日志

(55)

[ ]

[

= Eq (x1:T|x0)[日志p**e**(x0|x 1)] + Eq (x1:T|x0) 记录 + E q (x1:T|x0)

记录

]

 (56)

[ ]

[

+ E q (xt,xt — 1|x0) 记录

= Eq (x1|x0)[日志p**e**(x0|x1)] + Eq (xT|x0) 记录

 ] 

 (57)

T

= q (x1|x–0企)[ 志"p**e** (x| 1) - D ~~化~~、~~钾~~ (q(xT |x0 ) | p之(x~~T ))~~ - ~~Eq (xt|~~x0)[D氯化钾(q (xt - 1|xt,

reconstr执行期限 前ma刻度项 t=2 去噪匹配项

因此，我们成功地推导了ELBO的解释，可以估计较低，方差，因为每项被计算为一次最多一个随机变量的期望。 这个公式也有一个优雅的解释，这是在检查每个单独的术语时所揭示的：

1.Eq (x1|x0)[日志p**e**(x0|x 1)]可以解释为一个重建术语；就像它在普通VAE的ELBO中的类似物一样，这个项 可以使用蒙特卡罗估计进行近似和优化。

2.D氯化钾(q (xT|x0) | p (xT))表示最终噪声输入的分布与标准高斯先验的接近程度。它没有可训练的参数 , 在我们的假设下也等于零。

3.Eq(xt|x0)[D氯化钾(q (xt - 1|xt, x0) | p**e**(xt - 1|xt是一个去噪匹配项。我们学习了期望的去噪转 换步骤p**e**(xt - 1|xt)作为可处理的地面真实去噪过渡步骤q (xt - 1|xt, x0).q (xt - 1|xt, x0)过渡 步骤可以作为一个地面真实信号，因为它定义了如何去噪噪声图像xt可以访问最终的，完全去噪的图像x0 应该是因此，当两个去噪步骤尽可能接近时，这一项是最小化的，通过它们的KL发散度来测量。

作为附注，我们注意到，在两个ELBO推导（方程45和方程58）的过程中，只使用了马尔可夫假设；因此，这些公 式将适用于任何任意的马尔可夫HVAE。此外，当我们设置T = 1时，对VDM的两种ELBO解释都精确地重现了香草

VAE的ELBO方程，如方程式19所示。

在ELBO的推导中，优化代价的大部分再次是总和项，它主导了重建项。而每个KL的散度项D氯化钾(q (xt - 1|xt, x0) | p**e**(xt - 1|xt))在任意复杂的同时学习编码器中，任意的后验很难最小化，在VDM中，我们可以利用高

斯跃迁假设使优化易于处理。根据贝叶斯规则，我们有：

q (xt - 1|xt, x0)q t |xt -q1(x,t**x**|0x q (xt - 1 |x0 )

因为我们已经知道了，q (xt|xt - 1, x0) = q (xt|xt - 1) = N (xt ; ^ α t**x**t - 1, (1 - α t从我们关于编 码器转换的假设（方程31），剩下的是关于q (x的形式t|x0)和q (xt - 1|x0).幸运的是，通过利用VDM的编码

器转换是线性高斯模型这一事实，这些也使得它们易于处理。回想一下，在重新参数化技巧下，采样

xt~q (xt|xt - 1)可以重写为：

**x**t= ^ α t**x**t - 1+ ^ 1- α te与e N (e；0、I）

同样地，样本xt - 1~q (xt - 1|xt -2)可以重写为：

**x**t - 1= ^ α t - 1**x**t -2+^ 1 - α t - 1e与e N (e；0、I）

(59)

(60)

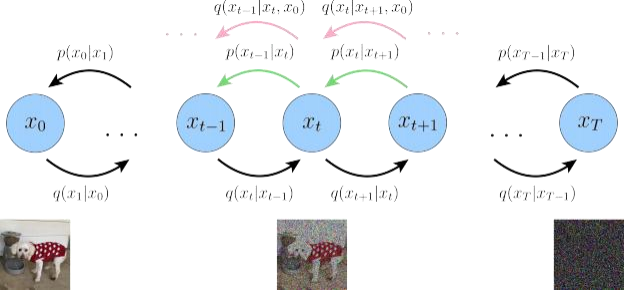


图5：描述的是另一种低方差的方法来优化VDM；我们计算地面真实去噪步骤q (x的形式t — 1|xt, x0)，利用 贝叶斯规则，用近似去噪步骤p最小化其KL散度**e**(xt — 1|xt).通过将绿色箭头表示的分布与粉色箭头表示，这 再次表示。艺术自由在这里发挥作用；在全貌中，每一个粉红色的箭头也必须来自x0 ，因为它也是一个条件反射 项。

然后是q的形式t|x0)可以通过重复应用重新参数化技巧来递归地推导出来。假设我们可以得到2T个随机噪声变量

{e，e t}

N(e ; 0, I).然后，对于一个任意的样本xtq (xt|x0)，我们可以将其改写为：

**x**t= ^ α t**x**t — 1+^ 1- α t**e** — 1

=^ α t (^ α t^ — 1**x**t — 2+^ 1 - α t — 1**e** —2)+^1-

α t**e** — 1= α tα t — ~~1~~**~~x~~**~~t~~ —~~2~~+^ α t- ~~α tα t~~  —

12 -—α~~tα~~2^t1tt 2 **x**t

α

^^

t-

—

2+ α t- α tα t — 1+ 1 - α t**e**t —2= α tαt —

1**x**t —2+^1 - α tαt — 1**e**t —2

= . . . 

'「t't 「

=/  α i**x**0+ / 1 - α i**e**0

= ^ t**x**0+ ^ 1 - t**e**0

N(xt ; ^ t**x**0, (1 - t) I)

(61) (62)

(63)

(64) (65) (66) (67)

(68)

(69) (70)

在方程64中，我们利用了两个独立的高斯随机变量的和仍然是一个高斯随机变量的事实 意思 是两个平均值的和 , 方差是 总数 的 两个方差。 口译^1- α t**e** — 1作为一个样本，从高斯N (0， (1 - αtI），和^ α t- α tα t — 1**e** —2作为高斯N (0， (αt- α tα t — 1)I)，然后我们可以把它们的和作为一个随机变量抽样来自高斯分 布的N (0， (1 - αt+ α t- α tα t — 1)I) = N(0, (1 - α t α t — 1 )I). 来自这个分布的样本可以然后 用重新参数化的技巧来表示为^1- α tα t — 1**e**t —2，如等式66所示。

因此，我们推导出了q (x的高斯形式t|x0).这个推导也可以被修改为得到描述q (x的高斯参数化t — 1|x0).现 在，了解了两个q的形式t|x0)和q (xt — 1|x0)，我们可以继续计算q (x的形式t — 1|xt, x0)，通过代入贝

叶斯规则展开：

q (xt |xt — 1 , **x**0 )q (xt — 1

q (xt — 1|xt, x0)| =

q (xt|x0

N(xt ; ^ α t **x**t — 1 , (1 - α t )I)N(xt — 1; ~~^ā~~t — 1 **x**0 , (1 - āt — 1 )I)

=

^ t**x**0 , ( 1 -t) I )

N(xt ;

~~1)2~~  —1

{ / exp ~~2~~]}

α ~~1~~

[ (x t tt

 —1  ~~)2~~

(xt ~~2~~

(xt

-t  )0)

**x** )0

**x**

α-t α-t

~~2~~  ( 1

+

αt)

—

— ~~)2~~ —

{

[

^ α

(xt  ~~tx~~ t

~~1~~  αt ]} { [

- —~~)2~~   — ~~1~~

+ (xt

1

= exp

α -t

~~1~~

( 2 αt**x** t **x**t— 1

-~~t~~ ~~)2~~

(xt ~~1~~ -

1 2

1

1)+ (x —

+ αt**x** — αt

= exp

1

~~t~~

[

]} { + C(xt, x0) / exp

+

]} { [

= exp

2

t —

+

( 1 2

( + )x

) ]} { [

αt  ~~1~~ α- t )—

**x**t — 1 = exp

~~1~~

)+ ~~1~~

(1 α-t —

+

( 1 2

~~(~~

{ exp [

α t [



)

]}

α-t)+ 1 α(t 1

α )t**x** —

+ **x**t — 1 =

αt

(1 ~~1~~ α) ~~t~~

)

~~1~~

) ]} {

(

**x** — 1

**x** —

+

2

t 1)

(1

**x**t — 1 = exp

—

) ]}

( 2

2 )

(

2

~~1~~

**x**t — 1 = exp

+

) 4**x**t — 1 2

1

(1

α) ~~t~~  t

—

3

x

t — 15

)2

(

4**x** — 1 2 x

~~1~~

(1 α) t t 1)

= exp

3

5

t —

t)x0**x**t

—

( ) [

{ exp

^ α 2

**x** — 1

t(1 α

1 =

]}

(71) (72) (73) (74) (75) (76) (77) (78) (79) (80)

(81)

(82)

(83)

(84)

N xt — - t — 1 (1 α ~~t~~ 0-(1 ~~t~~ 1)

uq (xx

对q (t)

在方程75中，C (xt, x0)是一个相对于x的常数项t — 1计算为只有x的组合t, x0 ，和α值；在等式84中隐式地 返回此项，以完成该平方。

因此，我们已经证明了在每一步中，xt — 1~q (xt — 1|xt, x0)为正态分布，均值为uq (xt, x0)，这是x的一 个函数t和x0，方差对q (t)作为α系数的函数。这些α系数在每个时间步长都是已知和固定的；当建模为超参数 时，它们要么被永久设置，要么被视为寻求建模它们的网络的当前推理输出。按照方程84，我们可以将我们的方

差方程重写为对q (t) = a (t)I，其中：

a(t) = ~~α)~~- ~~t~~— ~~)~~ (85)

为了匹配近似去噪转换步骤p**e**(xt — 1|xt)到地面真实去噪转换步骤q (xt — 1|xt, x0)尽可能接近，我们也 可以将其建模为高斯模型。此外， 由于已知所有的α项在每个时间步长都被冻结，我们可以立即构造近似去噪转

移步骤的方差也为对q (t) = a (t)I. 我们必须参数化它的均值u**e**(xt，t)作为x的函数t然而， 自从p**e**(xt —

1|xt)对x没有任何条件0.

回想一下，两个高斯分布之间的KL散度为：

D氯化钾(N(x ; ]

1 (uy- u北)

在我们的例子中，

u北, 对北) | N(y; uy, ~~对~~y~~))~~ =[记录y北 - d + tr (对1对北) + (uy- u北)T对

(86)

我们可以将两个高斯分布的方差精确匹配，优化KL散度项简化为最小化两个分布的平均值之间

的差异：

= arg最小D氯化钾(N(xt — 1; uq, 对q (t)) | N(xt — 1; u**e**, 对q (t))) **e**

= arg分 1 [记录 对 - d + tr (对 (t)— 1对 (t)) + (u**e**- u )T对 (t)— 1 (u**e**-

] **e** 2 对 q q q q

u q) 1

= arg2 [ 日1志1- d + d + (u**e**- u q)T对q (t)— 1 (u**e**- u q)]

= arg分 [ (u**e**- u q)T对q (t)— 1 (u**e**- u q)]

**e** 2

1[ ) 1 ]

= arg分 (u**e**- u q)T (a (t)I2 — (u**e**- u q)

**e** 2 [ ] q

(87)

(88)

(89)

(90)

(91)

(92)

我们给你写在了哪里q作为你的速记q (xt, x0)和u**e**作为你的速记**e**(xt，t)为简洁。换句话说，我们想优化一个 u**e**(xt，t)匹配你q (xt, x0)， 由我们导出的方程84可知，其形式为：

**u**q (xt, x0) =

αt (1 - t — 1 )xt +^ t — 1 (1 - α t )x0

1 - t

如u**e**(xt，t)对x也有条件t，我们可以匹配你q (xt, x0)，并将其紧密地设置为以下形式：

^ α t (1 - t — 1 )xt +^t — 1 (1 - αt )**e** (xt , t)

1 - t

**ue**(xt, t) =

在哪里2**e**(xt，t)是由一个试图预测x的神经网络参数化的0来自噪声图像xt和时间指数t。然后，将优化问题简 化为：

arg最小D氯化钾(q (xt — 1|xt, x0) | p**e**(xt — 1|xt))

**e**

= arg最小D氯化钾(N (xt — 1; uq, 对q (t)) | N (xt — 1; u**e**, 对q (t)))

(95)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **e**  = arg分  **e**  = arg分  **e**  = arg分  **e** |  | [  -  [  ] 2  -  [  2 ]  2  ( **e**(xt, t) - x0)  2 |  | ] 2  2 |

|  |  |  |  |
| --- | --- | --- | --- |
| = arg分  **e** |  | [ | ]  |**e**(xt, t) - x0| |

(96)

(97)

(98)

(99)

因此，优化VDM归结为学习神经网络，从任意噪声的版本[5]预测原始地面真实图像。此外，最小化我们导出的

ELBO目标（方程58）的求和项可以通过最小化所有时间步长的期望来近似：

[

(100)

U{2,T } Eq (xt|x0)[D氯化钾(q (xt — 1|xt, x0) | p**e**(xt —

arg最小E t 1|xt))]]

**e**

然后可以在时间步长上使用随机样本进行优化。

**学习扩散噪声参数**

让我们来研究如何共同学习一个VDM的噪声参数。一种可能的方法是对α进行建模t使用一个带有参数的神经网 络(t)。然而，这是低效的，因为推理必须在每个时间步t执行多次才能计算t.尽管缓存可以减轻这种计算成本 , 但我们也可以推导出另一种方法来学习扩散噪声参数。通过将方程85中的方差方程代入方程99中导出的每时间

步程目标，我们可以减少：

[ ]

[ ]

|**e**(xt, t) — x0| (101)

|**e**(xt, t) — x0| =

[ ]

|**e**(xt, t) — x0| (102)

=

[ ]

|**e**(xt, t) — x0| (103)

=

[ ]

|**e**(xt, t) — x0| (104)

=

[ ]

|**e**(xt, t) — x0| (105)

=

[ ]

|**e**(xt, t) — x0| (106)

=

) [ ]

(

=  — (

= 从方程70中回想一下，q (xt|x0)是一个形式为N (x的高斯分布t ; ^

|**e**(xt, t) — x0| (107)

—  |**e**(xt, t) — x0| (108)

) [ ]

t**x**0, (1 — t) I).然后，根据信噪比(

SNR）的定义为信噪比=，我们可以将每个时间步长t的信噪比写为：

t

1 —

SNR (t) =

然后，我们导出的方程108（和方程99)可以简化为：

|**e**(xt, t) — x0| = (SNR (t — 1) — SNR (t))

[ ]

]

x0|

[

|**e**(xt, t) —

(109)

(110)

顾名思义，信噪比表示原始信号与所存在的噪声量之间的比值；较高的信噪比代表越多的信号，较低的信噪比代 表越多的噪声。在扩散模型中，我们要求信噪比随着时间步长t的增加而单调地减小；这就形式化了扰动输入x的 概念t随着时间的推移，它变得越来越有噪声，直到它在t = T处变得与标准高斯分布相同。

根据方程110中目标的简化，我们可以使用神经网络直接参数化每个时间步长的信噪比，并将其与扩散模型共同 学习。 由于信噪比必须随时间单调下降，我们可以将其表示为：

SNR (t) = exp ( — ! (t)) (111)

在哪里(t)被建模为一个具有参数的单调递增的神经网络。关闭！(t)导致一个单调递减的函数，而指数强迫 得到的项是正的。请注意，方程100中的目标现在也必须进行优化。n通过将方程111中的信噪比参数化与方程109 中的信噪比定义相结合，我们也可以明确地推导出为的值的优雅形式t以及1的值t:

1 t = exp ( — ! (t))

) t=乙状结肠(-！(t))

) 1 — t=乙状体(！(t))

(112)

(113) (114)

这些术语对于各种计算都是必要的；例如，在优化过程中，它们被用来创建任意有噪声的xt从输入x0使用重新参 数化技巧，如等式69所示。

**三等效解释**

正如我们之前所证明的，一个变分扩散模型可以通过简单地学习一个神经网络来预测原始的自然图像x0从一个任 意的噪声版本xt和它的时间指数t。然而，x0有另外两个等价的参数化，这导致了对VDM的两个进一步的解释。

首先，我们可以利用重新参数化的技巧。在我们推导的q (x的形式中t|x0)，我们可以重新排列公式69来表明：

**x** 0= **x**t - ^  **e**0

把它插入我们之前推导的真去噪转换意味着uq (xt, x0)，我们可以重新推导为：

|  |  |
| --- | --- |
| **u** q (xt, x0)  =  =  = | α t (1 - t — 1 )xt + t — 1 (1 - α t )x0  ^  1 - t  ^ α t (1 - t — 1 )xt +  ~~^~~ t— 1 (1 - αt )  1 - t  α t (1 - t — 1 )xt + (1 - α t )  ^  1 -āt  +  ^ α t (1 - t — 1 )xt (1 - α t)xt(1 - α t)^1-~~t~~**~~e~~**~~0~~  1 - t(1 - t)^α t(1 - t)^ α t |

(  )

**~~x~~**~~t~~-

+

=

(

)

+

**x**t-

e e

0 =



t- e 0

0

x

x

x t- e

=

t- e 0

=

 0

=

因此，我们可以设置我们的近似去噪转移均值u**e**(xt，t），如：

**ue**(xt, t) = x t- **e**(xt, t)

然后，相应的优化问题变为：

arg最小D氯化钾(q (xt — 1|xt, x0) | p**e**(xt — 1|xt))

**e**

= arg最小D氯化钾(N (xt — 1; uq, 对q (t)) | N (xt — 1; u**e**, 对q (t)))

|  |  |
| --- | --- |
| **e**  =参数最小值 1  **e**2a(t)  =参数最小值 1  **e**2a(t)  =参数最小值 1  **e**2a(t) | [  0  ] [  2  2  **e**0-    **e**(xt, t)  **x**t-  t+  **e**(xt, t) - x  e  [  ]  2  **e**(xt, t))  (e0-  2 |

(115)

=

|  |  |
| --- | --- |
| ]  2  2 | a  r  g  分  **e**    [  |  e  0    **e**  (  x  t  , |

]

t)|

(116)

(117)

(118) (119) (120) (121) (122) (123) (124)

(125)

(126)

(127)

(128)

(129)

(130)

这里 **e**(xt； t)是一个学习预测源噪声e的神经网络0~确定x的N (e、0、I）t从x0.因此，我们已经证明了通过预 测原始图像x来学习一个VDM0相当于学习预测噪声；然而，根据经验，一些工作发现，预测噪声会导致更好的性 能[5,7]。

(xt)= +  **x**t+日志p (xt)=

x t+日志p (xt) =

x t+日志p (xt)

= x t+日志p (xt)

为了推导变分扩散模型的第三种常见解释，我们利用特威迪的公式[8]。在英语中，Tweedie的公式指出，一个指 数族分布的真实平均值，给定从它中抽取的样本，可以通过样本的最大似然估计（即经验平均值）加上一些涉及 估计的分数的修正项来估计。在只有一个观察样本的情况下，经验平均值只是样本本身。它通常用于减轻样本偏 差；如果观察到的样本都位于潜在分布的一端，则负分数变得很大，并将样本的朴素最大似然估计修正为真实均 值。

在数学上，对于一个高斯变量z N (z；uz ; 对z)，Tweedie的公式指出：

E [uz|z] = z + 对zr**z**日志p (z)

在这种情况下，我们用它来预测x的真实后验均值t考虑到它的样本。从方程70中，我们知道：

q (xt|x0) = N(xt ; ^ t**x**0; (1 - t) I)

然后，根据Tweedie的公式，我们有：

E [ux t|xt] = xt+ (1 -āt)r **x**t日志p (xt) (131)

我们写作的地方**x**t日志p (xt）作为r 日志p (xt)为名义上的简单。 根据公式Tweedie，最好的最佳估计是xt是由 ^

, ux t= t**x**0 ，定义为：

^

t**x**0= xt+ (1 - t)r 日志p (xt)

) x0= **x**t + (1 - ^ ~~t~~ )tr记录p(x~~t~~ )

然后，我们可以将方程133插入我们的地面真去噪转换均值u q (xt ; x0)，并再次推导出一个新的形式：

|  |  |
| --- | --- |
| **u**q (xt ; x0)  =  =  =  = | ^ ~~α~~ t (1 - t — 1 )xt +^ t — 1 (1 - α t )x0  1 - t  ^ α t(1 - t — 1)xt+^ t — 1(1 -  α t)  1 -āt  αt (1 - t — 1 )xt + (1 - α t )  ^  1 - t  + +  ^  α t (1 - t — 1 )xt (1 - α t)xt(1 - α t)(1 - t)r 日志p (xt)  1 - t(1 - t)^αt(1 - t)^ α t |

(  )

= +  **x**t+日志p

( )

(132)

(133)

(134)

(135)

(136) (137) (138) (139) (140) (141) (142)



因此，我们也可以设置我们的近似去噪转移均值u**e**(xt；t）为：

**ue**(xt ; t) = x t+ s **e**(xt ; t)

然后，相应的优化问题变为：

arg最小D氯化钾(q (xt — 1|xt ; x0) | p**e**(xt — 1|xt))

**e**

= arg最小D氯化钾(N (xt — 1; u q; 对q (t)) | N (xt — 1; u**e** ; 对q (t)))

=参数 [ **x**t+ s **e**(xt ; t) - x t -r 日志p

q ] [ ]

2 2

=参数最小值 1 (xt) **se**(xt ；t) - r 日志p (xt)

**e**2a(t) 2 2

[ ]

=参数最小值 1 (s**e**(xt；t) - r 日志p (xt))

2

**e**2a(t) 2

|  |  |  |
| --- | --- | --- |
| = arg分  **e** |  | [ ]  |s**e**(xt；t) - r 日志p (xt)| |

(143)

(144)

(145)

(146)

(147)

(148)

这里，s**e**(xt；t)是一个学习预测分数函数r的神经网络**x**t日志p (xt)，这是x的梯度t在数据空间中，对于任意 任意的噪声水平t。

精明的阅读器会注意到分数函数r log p (xt)在形式上与源噪声e非常相似0.这可以通过结合Tweedie的公式（公 式133）和重新参数化技巧（公式115）来明确地显示：

**x**0= = (149)

) (1 -āt)r 日志p (xt) = -^1-āt**e**0(150) r 日志(xt) = - e0(151)

事实证明，这两项被一个随时间变化的常数因子抵消了！分数函数测量如何在数据空间中移动以最大化对数概率 ; 直观地说， 由于源噪声被添加到自然图像中以破坏它，向相反的方向移动会“去噪 ”图像，这将是增加后续对 数概率的最佳更新。我们的数学证明证明了这种直觉；我们已经明确地表明，学习对分数函数进行建模等价于建 模源噪声的负值（高达一个比例因子）。

因此，我们推导出了三个等效的目标来优化一个VDM：学习一个神经网络来预测原始图像x0 ，源噪声e0 ，或图像 在任意噪声水平r log p (xt).VDM可以通过随机采样时间步长t和使用地面真实目标最小化预测的范数来进行可 伸缩性的训练。

**基于分数的生成模型**

我们已经证明了一个变分扩散模型可以简单地通过优化一个神经网络s来学习**e**(xt来预测分数函数r log p (xt). 然而，在我们的推导中，分数术语来自于Tweedie公式的应用；这并不一定为我们提供伟大的直觉或洞察分数函

数到底是什么或为什么它值得建模。幸运的是，我们可以寻找另一类生成模型，即基于分数的生成模型[9,10,11] , 来实现这种直觉。事实证明，我们可以证明，我们之前推导出的VDM公式有一个等价的基于分数的生成建模公

式，使我们可以随意在这两种解释之间灵活切换。

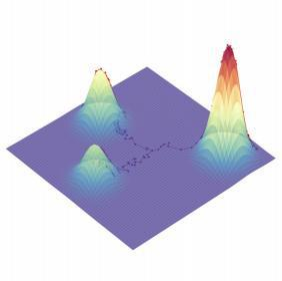
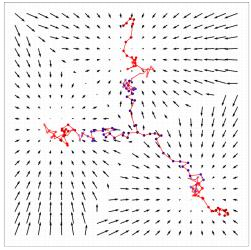


图6： 由朗之万动力学生成的三个随机采样轨迹的可视化，它们都从相同的初始化点开始，对于一个高斯混合物 。左图在三维轮廓上绘制这些采样轨迹，而右边的图根据地面真相得分函数绘制采样轨迹。从相同的初始化点， 由于朗之万动态抽样过程中的随机噪声项，我们能够从不同模式的样本；没有它，从一个固定的点抽样，每次试 验都总是确定地遵循分数到相同的模式。

为了理解为什么优化分数函数是有意义的，我们绕道而行，重新访问基于能量的模型[12,13]。任意灵活的概率 分布可以写成以下形式：

p**e**(x) = e — f**e**(x) (152)

其中f**e**(x)是一个任意灵活的，可参数化的函数，称为能量函数，通常由一个神经网络建模，和Z**e**是一个标准化 常数来确保的吗l p**e**(x) dx = 1. 学习这种分布的一种方法是最大似然法；然而，这需要精确地计算归一化常 数Z**e**= l e— f**e**(x) dx，这可能不是对复杂的f**e**(x)函数。

避免计算或建模归一化常数的一种方法是使用神经网络s**e**(x)学习分布p (x)的分数函数r log p (x)。这是由于 观察到，对公式152两边的对数的导数可以得到：

r**x**日志p**e**(x) = r **x**日志（e — f**e**(x))

= r **x**记录Z**e**+ r **x**日志e— f**e**(x )

= r **x**f**e** (x)

**se**(x)

(153)

(154)

(155) (156)

可以自由地表示为神经网络，不涉及任何归一化常数。分数模型可以通过使用地面真实分数函数最小化Fisher散 度来进行优化：

[ ]

Ep (x) |s**e**(x) r log p (x)| (157)

分数函数代表什么？对于每一个x，取它对x的对数似然的梯度，本质上描述了在数据空间中移动的方向，以进一 步增加其可能性。

然后，直观地说，分数函数在数据x所居住的整个空间上定义了一个向量场，指向模式。从视觉上看，这一点如 图6的右图所示。然后，通过学习真实数据分布的得分函数，我们可以从同一空间中的任意一点开始，迭代地跟 踪得分，直到达到一个模式。这种采样过程被称为朗之万动力学，在数学上被描述为：

**x**i+1**x**i+ cr 日志p (xi) +^2ce，我= 0, 1, ..., K (158)

其中x0从先验分布（如均匀） 中随机抽样，e N（e；0,0，I）是一个额外的噪声项，以确保生成的样本不总是崩 溃到一个模式，而是在其周围盘旋以获得多样性。~此外， 由于学习到的分数函数是确定性的，因此涉及到噪声 项的抽样增加了生成过程的随机性，允许我们避免确定性的轨迹。当从位于多个模式之间的位置初始化采样时， 这一点特别有用。朗之万动态采样和噪声项的好处的可视化描述如图6所示。

请注意，公式157中的目标依赖于对地面真实得分函数的访问，这对于复杂的分布(我们来说是不可用的。幸运的 是，在不知道地面真实分数的情况下，可以最小化Fisher散度，并且可以通过随机梯度下降进行优化。

总的来说，学习将一个分布表示为一个分数函数，并使用它通过马尔可夫链蒙特卡罗技术生成样本，如朗之万动 力学，被称为基于分数的生成建模[9,10,11]。

香草分数匹配有三个主要问题，如Song和Ermon [9]详细介绍的那样。首先，当x位于高维空间的低维流形上时， 分数函数的定义不明确。这可以从数学上看到；不是在低维流形上的所有点的概率都为零，其对数是未定义的。 当试图学习自然图像上的生成模型时，这尤其不方便，因为自然图像位于整个环境空间的低维流形上。

其次，通过香草分数匹配训练的估计分数函数在低密度区域是不准确的。这从我们在方程157中最小化的目标中 可以明显看出。因为它是在p (x)上的期望，并且明确地对其样本进行训练，模型将无法接收到很少看到或未看 到的例子的准确学习信号。这是有问题的，因为我们的采样策略涉及从高维空间中的一个随机位置开始，这很可 能是随机噪声，并根据学习到的分数函数移动。 由于我们遵循的是一个有噪声或不准确的分数估计，最终生成的 样本也可能是次优的，或者需要更多的迭代来收敛于一个准确的输出。

最后，朗之万动态抽样可能不会混合，即使它是使用地面真实分数执行的。假设真实的数据分布是两个不相交的 分布的混合物：

p (x) = c1p1(x) + c2p2(x) (159)

然后，当计算分数时，这些混合系数就会丢失，因为对数操作将系数从分布中分割出来，而梯度操作将其归零。 为了可视化这一点，请注意，右图6所示的地面真实得分函数不知道三个分布之间的不同权重；从描述的初始化 点采样的朗之万动力学的概率大致相等

尽管右下角的模式在实际的高斯混合中有更高的权重。

结果表明，通过在数据中添加多个层次的高斯噪声，可以同时解决这三个缺点。首先， 由于对高斯噪声分布的支 持是整个空间，一个被扰动的数据样本将不再局限于一个低维流形。其次，增加较大的高斯噪声会增加各模式在 数据分布中覆盖的面积，在低密度区域增加更多的训练信号。最后，随着方差的增加，增加多个层次的高斯噪声 将会得到尊重地面真实混合系数的中间分布。

在形式上，我们可以选择一个正的噪声水平序列{at}，并定义一系列逐步扰动的数据分布： \

pat(xt) = p (x)N(xt ; x, aI)dx (160)

然后，一个神经网络s**e**（x，t）采用分数匹配法进行学习，同时学习所有噪声水平的分数函数：

T

[ ]

λ(t)Ept(xt) |s**e**（x、t) - r log pat(xt)| (161)

arg分

**e**

其中， λ(t)是一个影响噪声水平t的正加权函数。请注意，这个目标几乎与方程148中导出的训练变分扩散模型 的目标完全匹配。此外，作者提出退火朗之万动力学采样作为生成程序，其中通过对每个t = T，T - 1，...，

2,1序列运行朗之万动力学产生样本。初始化是从一些固定的先验（如统一的）中选择的，每个后续的采样步骤 都从前一个模拟的最终样本开始。因为噪声水平随着时间步长t的增加而稳步下降，并且随着时间的推移我们使 步长减小，所以样本最终收敛到一个真正的模式。这直接类似于在变分扩散模型的马尔可夫HVAE解释中执行的采 样过程，其中一个随机初始化的数据向量在降低噪声水平上迭代细化。

因此，我们在变分扩散模型和基于分数的生成模型之间建立了明确的联系，包括其训练目标和抽样程序。

一个问题是如何自然地将扩散模型推广到无限个时间步长。在马尔可夫的HVAE视图下，这可以解释为将层次结构 的数量扩展到无穷大T！1 . 从基于等效分数的生成模型的角度来更清楚地表示这一点；在无限数量的噪声尺度下 , 图像在连续时间上的扰动可以表示为一个随机过程，因此可以用随机微分方程（SDE）来描述。然后通过反转 SDE进行采样，这自然需要估计每个连续值噪声水平[10]上的分数函数。SDE的不同参数化本质上描述了随时间推 移的不同扰动方案，使噪声过程[6]的灵活建模成为可能。

**指导**

到目前为止，我们只专注于对数据分布p (x)进行建模。然而，我们也经常对学习条件分布p（x|y）感兴趣，这 将使我们能够明确地控制我们通过条件反射信息y生成的数据。这形成了图像超分辨率模型的主干，如级联扩散 模型[18]，以及最先进的图像-文本模型，如DALL-E 2 [19]和Imagen [7]。

在每次迭代中，添加条件反射信息的一种自然方法只是与时间步长信息一起使用。回想一下公式32中我们的联合 分布：

T

p (x0:T) = p (xT) u p **e**(xt — 1|xt)

t=1

然后，为了将其转化为一个条件扩散模型，我们可以简单地在每个过渡步骤中添加任意的条件信息y，如下为：

T

p (x0:T|y) = p (xT) u p **e**(xt — 1|xt, y) (162)

t=1

例如，y可以是图像-文本生成中的文本编码，也可以是用来执行超分辨率的低分辨率图像。因此，我们能够像以

前一样，通过预测来学习VDM的核心神经网络 **e**(xt, t, y) x0,**e**(xt, t, y) e 0，或s**e**(xt，

loyxt|y)，为每个期望的解释和实现。

这个香草公式的一个警告是，以这种方式训练的条件扩散模型可能会学会忽略或淡化任何给定的条件反射信息。 因此，指导被提出作为一种更明确地控制模型给条件信息的权重的方法，以牺牲样本多样性为代价。两种最流行 的引导形式被称为分类器引导[10,20]和无分类器引导[21]。

**分类器指导**

让我们从一个基于分数的扩散模型的公式开始，其中我们的目标是学习r log p (xt|y)，条件模型的分数，在任 意噪声水平的t。回想一下，r是r的缩写**x**t为了简洁起见。通过贝叶斯规则，我们可以推导出以下等价形式：

( )

r 日志(xt|y) = 日志 

= 日志p (xt) + r 日志p (y |xt)r 日志p (y)

= 日志p (xt) + r 日志p (y |xt)

无件的企，对抗性的梯 "

企

物

(163)

(164) (165)

我们利用了log p (y)相对于x的梯度t为零。

我们最终得到的结果可以解释为学习了一个无条件的分数函数，并结合了一个分类器的对抗性梯度p (y |xt).因 此，在分类器引导[10,20]中，无条件扩散模型的分数与之前导出的一样，以及接受任意噪声xt并试图预测条件 信息y。然后，在采样过程中，将退火朗之万动力学的总体条件评分函数计算为无条件评分函数和噪声分类器的 对抗梯度的和。

为了引入细粒度控制来鼓励或阻止模型考虑条件反射信息，分类器引导通过一个V超参数项来衡量有噪声分类器 的对抗性梯度。在分类器指导下学习到的分数函数可以总结为：

r 日志(xt|y) = r 日志p (xt) + Vr 日志p (y |xt) (166)

直观地说，当V = 0时，条件扩散模型学会完全忽略条件信息，当V很大时，条件扩散模型学会产生严重遵守条件 信息的样本。这将以样本多样性为代价，因为它只会产生的数据很容易重新生成所提供的条件信息，即使是在嘈 杂的水平上。

分类器指导的一个显著缺点是它依赖于一个单独学习的分类器。 由于分类器必须处理任意噪声输入，而这是大多 数现有的预训练分类模型没有优化的，所以它必须与扩散模型一起特别学习。

**无分类器指导**

在无分类器指导[21]中，作者放弃了训练一个单独的分类器模型，转而选择无条件扩散模型和条件扩散模型。为 了推导出在无分类器指导下的分数函数，我们可以首先重新排列方程165，以表明：

r 日志p (y |xt) = r 日志p (xt|)rlogp (xt) (167)

然后，将它代入公式166，我们得到：

r 日志(xt|y) = r 日志p (xt) + V (r log p (xt|)rlogp (xt)) (168)

= 日志p (xt) + Vr 日志p (xt- |y) Vr 日志p (xt) (169)

=  r 日志p (xt |y) + (1  V)r 日志 p (xt) (170)

条件的企，无条件的分数 物 "

同样，V是一个控制我们所学到的条件模型对条件反射信息的关心程度的术语。当V = 0时，学习的条件模型完全 忽略调节器，学习无条件扩散模型。当V = 1时，模型在没有指导的情况下明确地学习了普通的条件分布。当V > 1时，扩散模型不仅对条件评分函数进行了优先排序，而且还朝着远离无条件评分函数的方向移动。换句话说，

它减少了生成不使用条件反射信息的样本的概率，而有利于明确使用条件反射信息的样本。这也有减少样本多样 性的效果，但代价是生成准确匹配条件信息的样本。

由于学习两个独立的扩散模型代价昂贵，我们可以将条件扩散模型和无条件扩散模型一起学习为一个奇异的条件 模型；无条件扩散模型可以通过用固定的常数值替换条件信息来查询，如零。这本质上是对条件反射信息执行随 机退出。无分类器的指导是优雅的，因为它使我们能够更好地控制我们的条件生成过程，而只需要只训练一个奇 异的扩散模型。

**关闭**

让我们在探索过程中总结我们的发现。首先，我们推导出变分扩散模型作为马尔可夫层次变分自编码器的一个特 殊情况，其中三个关键假设使ELBO的易于计算和可扩展优化成为可能。然后我们证明优化VDM归结为学习神经网 络预测三个潜在目标之一：原始源图像从任何任意噪声，任何任意噪声图像的原始源噪声，或噪声图像的得分函 数在任何任意噪声水平。然后，我们更深入地研究学习分数函数的含义，并明确地将其与基于分数的生成建模的 观点联系起来。最后，我们将讨论如何使用扩散模型来学习条件分布。

总之，扩散模型作为生成模型显示出令人难以置信的能力；事实上，它们为目前最先进的文本条件图像生成模型 提供了动力，如Imagen和DALL-E 2。此外，使这些模型成为可能的数学方法也非常优雅。然而，仍有一些缺点需 要考虑：

这不太可能是我们作为人类自然地建模和生成数据的方式；我们不会将样本作为随机噪声进行迭代去噪。

VDM不产生可解释的延迟。VAE希望通过优化其编码器来学习一个结构化的潜在空间，而在VDM中，每个时间步 长的编码器已经作为一个线性高斯模型给出，不能灵活地优化。因此，中间延迟被限制为原始输入的有噪 声版本。

延迟被限制在与原始输入相同的维度上，进一步阻碍了学习有意义的、压缩的潜在结构的努力。

采样是一个昂贵的程序，因为多个去噪步骤必须运行在两个公式下。回想一下，其中一个限制是选择了足够 多的时间步长T，以确保最终的潜在变量是完全高斯噪声；在采样过程中，我们必须遍历所有这些时间步长 来生成一个样本。

最后，扩散模型的成功突出了分层VAEs作为生成模型的力量。我们已经证明，当我们推广到inhnite潜在层次结 构时，即使编码器是平凡的，潜在维数是固定的，并且假设了马尔可夫转换，我们仍然能够学习到强大的数据模 型。这表明，在一般的、深度HVAEs的情况下，可以实现进一步的性能提高，其中可以潜在地学习复杂的编码器 和语义上有意义的潜在空间。

**感谢：我想感谢狄龙、杨松、金马、本普尔、何亚生、姜丹、陈婷、陈孙对该作品的草稿的审阅，并提供了许多** **有益的编辑和评论。**非常感谢！