使用 RealNVP 进行密度估计

Laurent Dinh
Montreal Institute for Learning Algorithms
University of Montreal
Montreal, QC H3T1J4

Jascha Sohl-Dickstein
Google Brain

Samy Bengio
Google Brain

摘 要

概率模型的无监督学习是机器学习中的一个核心但具有挑战性的问题。 具体而言，设计具有易学习，采样，推理和评估的模型对于解决此任务至关重要。 我们使用实值非体积保持（Real NVP）变换——一组强大，稳定，可逆和可学习的变换来扩展此类模型的空间，从而产生具有精确对数似然计算的无监督学习算法，精确且有效的采样，潜在变量的精确和有效推断，以及可解释的潜在空间。 我们展示了它通过采样，对数似然评估和潜变量操作在四个数据集上建模自然图像的能力。

1 介绍

由于改进的监督学习技术，表示学习领域经历了巨大的进步。然而，无监督学习有可能利用大量未标记数据，并将这些进展扩展到其他方式不切实际或不可能的地方。

无监督学习的一种原则性方法是生成概率建模。生成概率模型不仅具有创造新内容的能力，它们还具有广泛的重建相关应用，包括修复[61,46,59]，去噪[3]，着色[71]和超分辨率[9]]。

由于感兴趣的数据通常是高维度和高度结构化的，因此该领域的挑战是建立足够强大的模型来捕捉其复杂特征并且模型仍然可训练。我们通过引入实值非体积保持（Real NVP）变换来解决这一挑战，这是一种易于处理且富有表现力的高维数据建模方法。

该模型可以执行数据点的有效且精确的推断，采样和对数密度估计。此外，本文介绍的体系结构可以从该模型提取的分层特征中精确有效地重建输入图像。

2 相关工作

关于概率生成模型的大量工作侧重于使用最大似然训练模型。 一类最大似然模型是由概率无向图描述的模型，例如受限玻尔兹曼机器[58]和深玻波兹曼机器[53]。 通过利用其二分结构的条件独立性来训练这些模型，以允许对潜在变量进行有效的精确或近似后验推断。 然而，由于相关边际分布对潜在变量的难以处理，其训练，评估和采样程序需要使用近似值，如平均场推理和马尔可夫链蒙特卡罗，这些复杂模型的收敛时间

当作者在 Google Brain 时，工作已经完成.

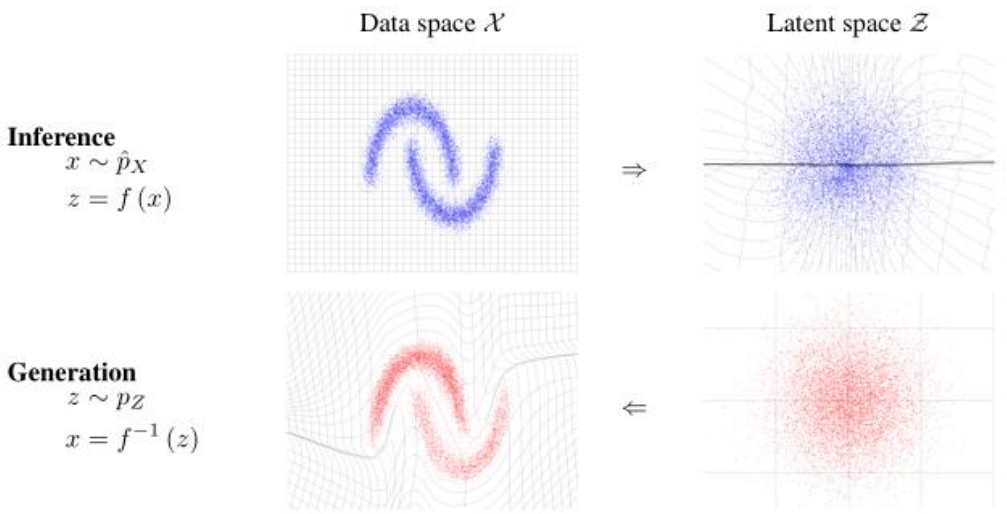


图 1：Real NVP 学习数据分布 $\hat{p}_X$ 和潜在分布 Pz（通常是高斯分布）之间的可逆，稳定映射。 在这里，我们展示了在玩具 2-d 数据集上学习的映射。 函数 f（x）将样本 x 从左上方的数据分布映射到右上方的潜在分布的近似样本 z。 这对应于给定数据的潜在状态的精确推断。 反函数$f^{-1}(z)$将样本 z 从右下方的潜在分布映射到左下方数据分布的近似样本 x。 这对应于从模型中精确生成样本。 对于 f（x）和$f^{-1}(z)$，图中还示出了 X 和 Z 空间中的网格线的变换。

仍未确定，通常导致产生高度相关的样品。此外，这些近似值通常会妨碍它们的性能[7]。

相反，定向图形模型是根据原始抽样程序定义的，这对于其概念和计算简单性都具有吸引力。然而，它们缺乏无向模型的条件独立结构，使潜在变量的精确和近似后验推断变得繁琐[56]。随机变分推理[27]和摊销推理[13,43,35,49]的最新进展，通过最大化对数似然的变分下界，允许有效的近似推理和深度定向图形模型的学习[45]。特别地，变自动编码算法[35，49]同时学习一个生成网络，映射 Z 到样本 x 高斯潜变量，并且通过利用重新参数化匹配的近似推理网络技巧[68]，将样本 x 映射到一个语义上有意义的潜表示 Z，。它成功地利用了深度神经网络中反向传播的最新进展[51,39]，从而将其应用于从语音合成[12]到语言建模[8]的多种应用。尽管如此，推理过程中的近似限制了其学习高维深度表示的能力，激发了近期改进近似推理的工作[42,48,55,63,10,59,34]。

通过避免使用潜在变量可以完全避免这种近似。自回归模型[18,6,37,20]可以实现这一策略，同时通常保留很大的灵活性。这类算法通过根据维度上的固定排序使用概率链规则将其分解为条件的乘积，从而简化对数似然评估和采样，从而可行地模拟联合分布。最近在这一系列研究中的工作利用了复发网络[51]的最新进展，特别是长期短期记忆[26]和剩余网络[25,24]，以便学习最新技术生成图像模型[61,46]和语言模型[32]。尺寸的排序虽然经常是任意的，但对模型的训练至关重要[66]。该模型的顺序性质限制了其计算效率，例如，它的采样过程是顺序的和不可并行化的，这在语音和音乐合成或实时渲染等应用中会变得很麻烦。此外，没有与自回归模型相关的自然潜在表示，它们还没有被证明对半监督学习很有用。

另一方面，生成性对抗网络（GAN）[21]可以通过完全避免最大似然原则来训练任何可微分的生成网络。相反，生成网络与鉴别器网络相关联，鉴别器网络的任务是区分样本和真实数据。该鉴别器网络不是使用难以处理的对数似然，而是以对抗的方式提供训练信号。成功训练的 GAN 模型[21,15,47]可以始终如一地生成清晰逼真的样本[38]。然而，衡量生成样本多样性的指标目前难以处理[62,22,30]。此外，他们的训练过程中的不稳定性[47]需要仔细的超参数调整以避免不同的问题。

训练将潜在变量 z ~ P_Z 映射到样本 x ~ P_X 的生成网络 g 理论上不需要像 GAN 那样的鉴别器网络，或者如变分自动编码器中的近似推断。实际上，如果 g 是双射的，则可以使用变量公式的变化通过最大似然来训练它:

$$p_X(x) = p_Z(z) \left| \det \left(\frac{\partial g(z)}{\partial z^T} \right) \right|^{-1}. \tag{1}$$

这个公式已经在几篇论文中讨论过，包括独立成分分析的最大似然公式（ICA）[4,28]，高斯化[14,11]和深度密度模型[5,50,17,3]。 由于非线性 ICA 解的存在性证明[29]表明，自回归模型可以被视为最大似然非线性 ICA 的易处理实例，其中残差对应于独立分量。 然而，变量公式的变化的不成熟应用产生了计算上昂贵且条件差的模型，因此这种类型的大规模模型尚未进入一般用途。

3 模型定义

在本文中，我们将通过最大似然性来解决在高维连续空间中学习高度非线性模型的问题。 为了优化对数似然，我们引入了一类更灵活的体系结构，可以使用变量公式的变化计算连续数据的对数似然。 基于我们之前在[17]中的工作，我们定义了一类强大的双射函数，可以进行精确和易处理的密度评估以及精确和易处理的推理。 此外，所得的成本函数不依赖于诸如平方误差[38,47]的固定形式的重建成本，并且能因此产生更清晰的样本。 此外，这种灵活性有助于我们利用批量归一化[31]和残差网络[24,25]的最新进展来定义具有多个抽象级别的非常深入的多尺度架构。

3.1 变量公式的变化

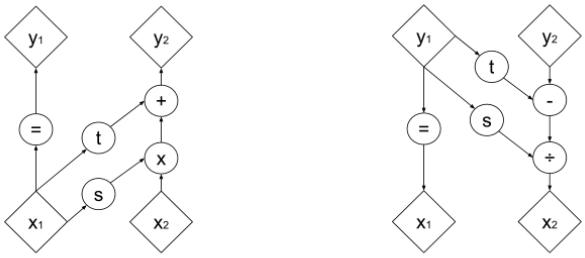
给定观测数据变量 $x \in X$，潜在变量 $z \in Z$ 上的简单先验概率分布 Pz 和双射 $f : X \to Z$ (with $g = f^{-1}$)，变量公式的变化定义 X 乘以的模型分布

$$p_X(x) = p_Z\big(f(x)\big) \left| \det \left(\frac{\partial f(x)}{\partial x^T} \right) \right| \tag{2}$$

$$\log\big(p_X(x)\big) = \log\Big(p_Z\big(f(x)\big)\Big) + \log\left(\left| \det \left(\frac{\partial f(x)}{\partial x^T} \right) \right| \right), \tag{3}$$

其中 $\frac{\partial f(x)}{\partial x^T}$ 是 x 的雅可比行列式。

可以使用逆变换采样规则[16]生成来自结果分布的精确样本。 在潜在空间中绘制样本 z ~ Pz，并且其反像 $x = f^{-1}(z) = g(z)$ 在原始空间中生成样本。 计算点 x 上的密度是通过计算其图像的密度 f（x）并乘以相关的雅可比行列式 $\left(\frac{\partial f(x)}{\partial x^T}\right)$ 来实现的。另请参见图 1。精确有效的推理可以准确，快速地评估模型。



(a) Forward propagation (b) Inverse propagation

图 2：正向和反向传播的计算图。 耦合层应用简单的可逆变换，该变换包括缩放，然后向输入向量 X1 的剩余部分上调节的输入向量的一部分 X2 添加恒定偏移。 由于其简单的性质，这种转变既易于翻转，又具有易处理的行列式。 然而，由函数 s 和 t 捕获的这种变换的条件性质显著地增加了这种其他弱函数的灵活性。 前向和反向传播操作具有相同的计算成本。

3.2 耦合层

计算具有高维域和密码域的函数的雅可比行列式以及大矩阵的行列式通常在计算上非常昂贵。这与对双射函数的限制结合起来，使得等式 2 看起来对于任意分布的建模是不切实际的。

如[17]中所示，通过仔细设计函数 f，可以学习一种既易于处理又极其灵活的双射模型。正如计算变换的雅可比行列式对于应用该原理去有效地训练是至关重要的，这项工作利用了一个简单的观察，即三角矩阵的行列式可以作为其对角项的乘积有效地计算。

我们将通过堆叠一系列简单的双射来构建灵活且易处理的双射函数。在每个简单的双射中，输入矢量的一部分使用易于反转的函数来更新，但是它以复杂的方式依赖于输入矢量的剩余部分。我们将这些简单的双射中的每一个称为仿射耦合层。给定 D 维输入 x 和 d <D，仿射耦合层的输出 y 遵循等式

$$y_{1:d} = x_{1:d} \tag{4}$$

$$y_{d+1:D} = x_{d+1:D} \odot \exp\big(s(x_{1:d})\big) + t(x_{1:d}), \tag{5}$$

其中 s 和 t 代表规模和翻译，函数是从$R^d \mapsto R^{D-d}$的，$\odot$ 是 Hadamard 乘积或按元素乘积（见图 2（a））

3.3 属性

雅可比行列的这种转换是

$$\frac{\partial y}{\partial x^T} = \begin{bmatrix} \mathbb{I}_d & 0 \\ \frac{\partial y_{d+1:D}}{\partial x_{1:d}^T} & \operatorname{diag}\left(\exp\left[s\left(x_{1:d}\right)\right]\right) \end{bmatrix}, \tag{6}$$

其中$\operatorname{diag}\left(\exp\left[s\left(x_{1:d}\right)\right]\right)$是对角矩阵，其对角元素对应于向量$\exp\left[s\left(x_{1:d}\right)\right]$。 鉴于观察到雅可比是三角形，我们可以有效地计算其决定式$\exp\left[\sum_j s\left(x_{1:d}\right)_j\right]$。因为计算耦合的雅可比行列式层操作不涉及计算 s 或 t 的雅可比行列式，这些函数可以任意复杂。我们将使它们成为深度卷积神经网络。请注意，s 和 t 的隐藏层可以具有比其输入和输出层更多的功能。

在定义概率模型的背景下，这些耦合层的另一个有趣特性是它们的可逆性。 实际上，计算逆不比前向传播复杂

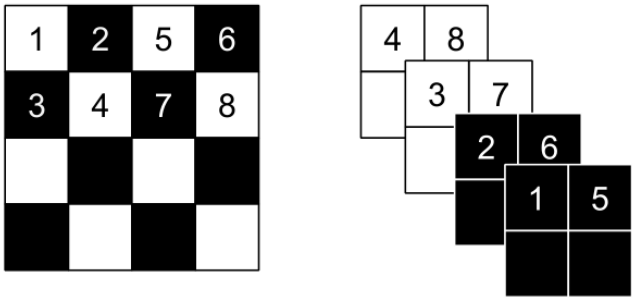


图 3：仿射耦合层的掩蔽方案。 左侧是空间棋盘图案蒙版。 在右边，一个通道扩大的掩蔽。 挤压操作将 4×4×1 张量（左侧）减小为 2×2×4 张量（右侧）。 在挤压操作之前，棋盘图案用于耦合层，而之后使用通道方式的掩模图案. (见图 2(b)),

$$\begin{cases} y_{1:d} & = x_{1:d} \\ y_{d+1:D} & = x_{d+1:D} \odot \exp\big(s(x_{1:d})\big) + t(x_{1:d}) \end{cases} \tag{7}$$

$$\Leftrightarrow \begin{cases} x_{1:d} & = y_{1:d} \\ x_{d+1:D} & = \big(y_{d+1:D} - t(y_{1:d})\big) \odot \exp\big(-s(y_{1:d})\big), \end{cases} \tag{8}$$

意味着采样与该模型的推断一样有效。 再次注意，计算耦合层的倒数不需要计算 s 或 t 的倒数，因此这些函数可能是任意复杂的并且难以反转。

3.4 掩码卷积

可以使用二进制掩码 b 并使用 y 的函数形式来实现分区，

$$y = b \odot x + (1 - b) \odot \Big(x \odot \exp\big(s(b \odot x)\big) + t(b \odot x)\Big). \tag{9}$$

我们使用两个分区来利用图像的局部相关结构：空间棋盘图案和通道屏蔽（参见图 3）。 空间棋盘图案掩模具有值 1，其中空间坐标的总和是奇数，否则为 0。 通道尺寸掩模 b 对于通道尺寸的前半部分是 1 而对于后半部分是 0。 对于此处提供的模型，s（.）和 t（.）都是经过整流的卷积网络.

3.5 结合耦合层

尽管耦合层可以很强大，但它们的正向变换使一些组件保持不变。 通过以交替模式组成耦合层可以克服这种困难，使得在一个耦合层中保持不变的组件在下一个中更新（参见图 4（a））。

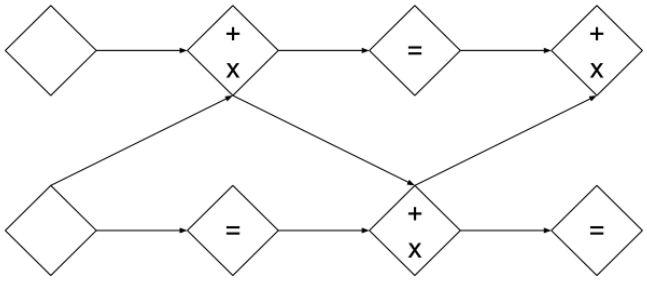
由此产生的函数的雅可比行列式仍然易于处理，考虑到

$$\frac{\partial(f_b \circ f_a)}{\partial x_a^T}(x_a) = \frac{\partial f_a}{\partial x_a^T}(x_a) \cdot \frac{\partial f_b}{\partial x_b^T}\big(x_b = f_a(x_a)\big) \tag{10}$$

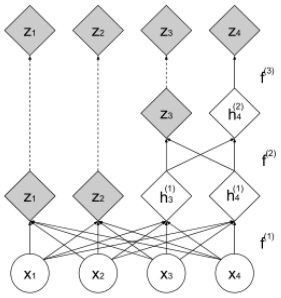
$$\det(A \cdot B) = \det(A)\det(B). \tag{11}$$

类似地，它的逆可以很容易地计算

$$(f_b \circ f_a)^{-1} = f_a^{-1} \circ f_b^{-1}. \tag{12}$$



（a）在这种交替模式中，在一个变换中保持相同的单元在下一个中被修改



（b）将变量分解。在每一步中，一半变量直接建模为高斯，而另一半则进一步变换。

图 4：仿射耦合层的组成方案

3.6 多尺度架构

我们使用压缩操作实现多尺度架构：对于每个通道，它将图像划分为形状为 2×2×c 的子方形，然后将它们重新整形为形状为 1×1×4c 的子方形。压缩操作将 s×s×c 张量转换为 s/2×s/2×4c 张量（见图 3），有效地交换通道数量的空间大小。

在每个尺度上，我们将几个操作组合成一个序列：我们首先应用三个具有交替棋盘面板的耦合层，然后执行挤压操作，最后应用三个具有交替通道屏蔽的耦合层。选择通道屏蔽以使得所得到的分区与先前的棋盘屏蔽不是多余的（参见图 3）。对于最终比例，我们仅应用具有交替棋盘面罩的四个耦合层。

在计算和存储器成本方面以及需要训练的参数数量方面，将 D 维向量传播通过所有耦合层是麻烦的。出于这个原因，我们遵循[57]的设计选择，并以规则的间隔分解出一半的尺寸（见公式 14）。我们可以递归地定义这个操作（见图 4（b））。

$$h^{(0)} = x \tag{13}$$

$$(z^{(i+1)}, h^{(i+1)}) = f^{(i+1)}(h^{(i)}) \tag{14}$$

$$z^{(L)} = f^{(L)}(h^{(L-1)}) \tag{15}$$

$$z = (z^{(1)}, \ldots, z^{(L)}). \tag{16}$$

在我们的实验中，我们对 i <L 使用该操作。当计算 f（i）时，每层执行上述耦合 - 压缩 - 耦合操作的序列（等式 14）。在每一层，随着空间分辨率的降低，s 和 t 中隐藏层特征的数量加倍。将在不同尺度上分解的所有变量连接起来以获得最终的变换输出（等式 16）。

因此，该模型必须对在较粗的尺度（在后面的层中）中分解出的那些单元进行高斯化，这些单元在更精细的尺度（在较早的层中）被分解出来。这导致了中间层次的表示[53,59]的定义，对应于附录 D 中所示的更局部，细粒度的特征。

此外，早期层中的高斯化和分解单元具有在整个网络中分配损失函数的实际益处，遵循类似于使用中间分类器引导中间层的原理[40]。它还显著减少了模型使用的计算量和内存量，使我们能够训练更大的模型。

3.7 批量归一化

为了进一步改善训练信号的传播，我们在 s 和 t 中使用深度残差网络[24,25]和批量归一化[31]和权重归一化[2,54]。 如附录 E 中所述，我们引入并使用了一种新的批量归一化变体，它基于最近的 minibatches 中的运行平均值，因此在使用非常小的 minibatches 进行训练时更加稳健。

我们还对整个耦合层输出使用批量归一化。 批量归一化的效果很容易包含在雅可比计算中，因为它在每个维度上充当线性重新缩放。 也就是说，给定估计的批次统计量 $\tilde{\mu}$ 和 $\tilde{\sigma}^2$，重新缩放函数

$$x \mapsto \frac{x - \tilde{\mu}}{\sqrt{\tilde{\sigma}^2 + \epsilon}} \tag{17}$$

有一个雅可比行列式

$$\left(\prod_i (\tilde{\sigma}_i^2 + \epsilon) \right)^{-\frac{1}{2}}. \tag{18}$$

这种形式的批量归一化可以被视为类似于深度强化学习中的奖励标准化[44,45]。

我们发现这种技术的使用不仅允许训练更深层次的耦合层，而且还减轻了训练者在通过基于梯度的方法训练条件分布时通常遇到的不稳定性问题.

4 实验

4.1 过程

公式 2 中描述的算法显示了如何学习无界空间上的分布。一般而言，感兴趣的数据具有有限的幅度。例如，图像的像素值通常位于 $[0, 256]^D$ 在应用推荐的抖动程序后[64,62]。为了减少边界效应的影响，我们改为模拟 logit 的密度 $(\alpha + (1-\alpha) \odot \frac{x}{256})$，其中α在这里选择.05。我们在计算对数似然和每维度位时考虑了这种转换。我们还在训练期间增加了 CIFAR-10，CelebA 和 LSUN 数据集，还包括训练样例的水平翻转。

我们在四个自然图像数据集上训练我们的模型：CIFAR-10 [36]，Imagenet [52]，大规模场景理解（LSUN）[70]，CelebFaces 属性（CelebA）[41]。更具体地说，我们训练下采样到 32×32 和 64×64 版本的 Imagenet [46]。对于 LSUN 数据集，我们在卧室，塔楼和教堂户外类别上进行训练。 LSUN 的程序与[47]中的相同：我们对图像进行下采样，使最小边为 96 像素，随机作物为 64×64。对于 CelebA，我们使用与[38]中相同的程序：我们采用大约 148×148 的中央作物然后将其调整为 64×64。

我们使用第 3.6 节中描述的多尺度架构，并在耦合层中使用深度卷积残差网络，如[46]所建议的整流器非线性和跳跃连接。为了计算缩放函数 s，我们使用双曲正切函数乘以学习尺度，而平移函数 t 具有仿射输出。我们的多尺度架构以递归方式重复，直到最后一次递归的输入为 4×4×c 张量。对于大小为 32×32 的图像的数据集，我们使用具有 32 个隐藏特征映射的 4 个残余块用于具有棋盘掩蔽的第一耦合层。只有 2 个剩余块用于 64×64 的图像。我们使用 64 的批量大小。对于 CIFAR-10，我们只使用 8 个残差块，64 个特征映射和仅缩减一次。我们使用默认的超参数 ADAM [33]进行优化，并在系数为 5*10-5 的体重秤参数上使用 L2 正则化。

我们将先前的 Pz 设置为各向同性单位范数高斯。但是，任何分布都可以用于 Pz，包括在训练期间也学习的分布，例如来自自回归模型，或者（稍微修改训练目标）变分自动编码器。

Dataset	PixelRNN [46]	Real NVP	Conv DRAW [22]	IAF-VAE [34]
CIFAR-10	3.00	3.49	< 3.59	< 3.28
Imagenet (32 32)	3.86 (3.83)	4.28 (4.26)	< 4.40 (4.35)	
Imagenet (64 64)	3.63 (3.57)	3.98 (3.75)	< 4.10 (4.04)	
LSUN (bedroom)		2.72 (2.70)		
LSUN (tower)		2.81 (2.78)		
LSUN (church outdoor)		3.08 (2.94)		
CelebA		3.02 (2.97)		

表 1：CIFAR-10，Imagenet，LSUN 数据集和 CelebA 的比特/暗淡结果。 CIFAR-10 的测试结果和 Imagenet，LSUN 和 CelebA 的验证结果（训练结果在括号中供参考）。

图 5：左侧列中的数据集示例。 在右栏中，来自模型的样本在数据集上训练。 此图中显示的数据集按顺序排列：CIFAR-10，Imagenet（32×32），Imagenet（64×64），CelebA，LSUN（卧室）。

4.2 结果

我们在表 1 中显示，每个维度的比特数虽然没有超过 Pixel RNN [46]基线，但与其他生成方法相比具有竞争力。 我们注意到我们的性能随着参数的数量而增加，更大的模型可能会进一步提高性能。 对于 CelebA 和 LSUN，验证集的每维度位数在整个训练期间都在减少，因此预计几乎不会过度拟合。

我们在图 5 中显示了从模型生成的样本，其中来自数据集的训练样例用于比较。 如[62,22]所述，最大似然是一种重视多样性的原则

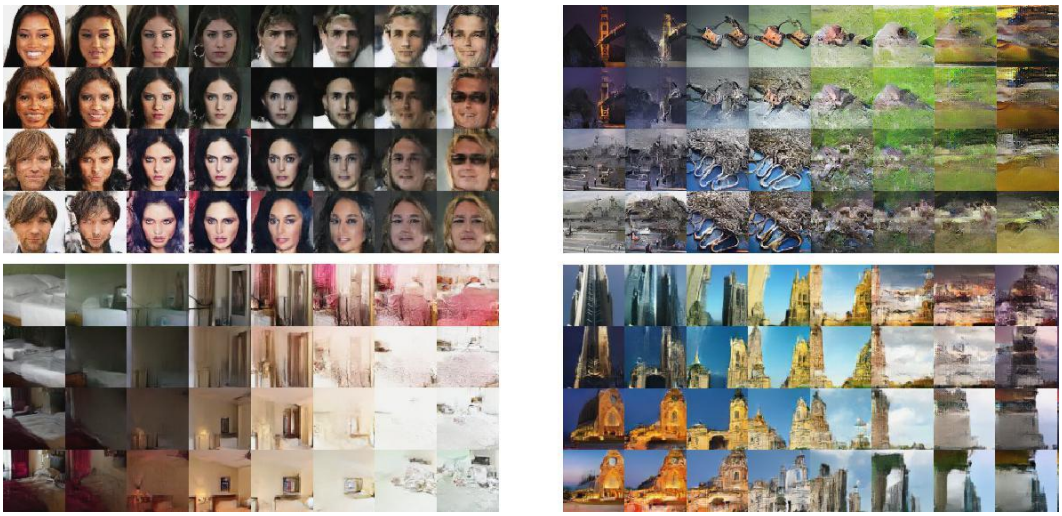


图 6：从数据集中的四个示例生成的流形。 从左上方顺时针方向：CelebA，Imagenet（64×64），LSUN（塔），LSUN（卧室）。

在有限的容量设置中超过样品质量。因此，我们的模型输出有时非常不可能的样本，我们可以特别注意到 CelebA。与变分自动编码器相反，从我们的模型生成的样本不仅看起来全局相干而且看起来很清晰。我们的假设是，与这些模型相反，Real NVP 不依赖于固定形式重建成本，如 L2 规范，它倾向于比高频分量更看重奖励捕获低频分量。与自回归模型不同，我们的模型采样非常有效，因为它在输入维度上并行化。在 Imagenet 和 LSUN 上，我们的模型似乎很好地捕捉了背景/前景和光照相互作用的概念，例如光度和反射和阴影的一致光源方向。

我们还说明了潜在变量的平滑语义一致意义。在潜在空间中，我们基于四个验证示例 z（1），z（2），z（3），z（4）定义了一个流形，并通过两个参数 φ 和 ϕ' 来参数化。

$$z = \cos(\phi)\left(\cos(\phi')z_{(1)} + \sin(\phi')z_{(2)}\right) + \sin(\phi)\left(\cos(\phi')z_{(3)} + \sin(\phi')z_{(4)}\right). \tag{19}$$

我们通过计算 g（z）将得到的流形投影回数据空间。 结果显示在图 6 中。我们观察到模型似乎已经组织了潜在空间，其意义概念远远超出了像素空间插值。 更多可视化显示在附录中。 为了进一步测试潜在空间是否具有一致的语义解释，我们在 CelebA 上训练了一个类条件模型，并发现学习表示在类标签上具有一致的语义（见附录 F）。

5 讨论和结论

在本文中，我们已经定义了一类具有易处理的雅可比行列式的可逆函数，能够进行精确且易处理的对数似然评估，推理和抽样。我们已经证明，这类生成模型在样本质量和对数似然方面都达到了竞争性。存在许多途径来进一步改进变换的功能形式，例如通过利用扩张卷积[69]和残差网络架构[60]的最新进展。

本文介绍了一种弥合自回归模型，变分自动编码器和生成对抗网络之间差距的技术。与自回归模型一样，它允许对训练进行易处理且精确的对数似然评估。然而，它允许更灵活的功能形式，类似于变分自动编码器的生成模型。这允许从模型分布中快速且精确地采样。与 GAN 一样，与变分自动编码器不同，我们的技术不需要使用固定形式的重建成本，而是根据更高级别的特征定义成本，从而生成更清晰的图像。最后，不同于变数

自动编码器和 GAN，我们的技术能够学习一个语义上有意义的潜在空间，它与输入空间一样高。这可能使算法特别适合半监督学习任务，因为我们希望在未来的工作中探索。

Real NVP 生成模型还可以以附加变量（例如类标签）为条件来创建结构化输出算法。更重要的是，由于所得到的可逆变换类可以被视为模块化方式的概率分布，它还可以用于改进其他概率模型，如自回归模型和变分自动编码器。对于变分自动编码器，这些变换既可用于实现更灵活的重建成本[38]，也可用于更灵活的随机推理分布[48]。概率模型通常也可以受益于本文中应用的批量归一化技术。

强大和可训练的可逆函数的定义也可以使生成无监督学习以外的领域受益。例如，在强化学习中，这些可逆函数可以帮助扩展 argmax 操作易于连续 Q 学习的函数集[23]或找到局部线性高斯近似更合适的表示[67]。

6 致谢

作者感谢 Tensorflow 的开发人员[1]。 我们感谢 Sherry Moore，David Andersen 和 Jon Shlens 帮助他们实施该模型。 我们感谢 Aäronvanden Oord，Yann Dauphin，Kyle Kastner，Chelsea Finn，Maithra Raghu，David Warde-Farley，Daniel Jiwoong Im 和 Oriol Vinyals 进行了富有成效的讨论。 最后，我们感谢 Ben Poole，Rafal Jozefowicz 和 George Dahl 对本文草案的意见。

参考文献

[1] Martın Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467, 2016.

[2] Vijay Badrinarayanan, Bamdev Mishra, and Roberto Cipolla. Understanding symmetries in deep networks. arXiv preprint arXiv:1511.01029, 2015.

[3] Johannes Ballé, Valero Laparra, and Eero P Simoncelli. Density modeling of images using a generalized normalization transformation. arXiv preprint arXiv:1511.06281, 2015.

[4] Anthony J Bell and Terrence J Sejnowski. An information-maximization approach to blind separation and blind deconvolution. Neural computation, 7(6):1129–1159, 1995.

[5] Yoshua Bengio. Artificial neural networks and their application to sequence recognition. 1991.

[6] Yoshua Bengio and Samy Bengio. Modeling high-dimensional discrete data with multi-layer neural networks. In NIPS, volume 99, pages 400–406, 1999.

[7] Mathias Berglund and Tapani Raiko. Stochastic gradient estimate variance in contrastive divergence and persistent contrastive divergence. arXiv preprint arXiv:1312.6002, 2013.

[8] Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. Generating sentences from a continuous space. arXiv preprint arXiv:1511.06349, 2015.

[9] Joan Bruna, Pablo Sprechmann, and Yann LeCun. Super-resolution with deep convolutional sufficient statistics. arXiv preprint arXiv:1511.05666, 2015.

[10] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. arXiv preprint arXiv:1509.00519, 2015.

[11] Scott Shaobing Chen and Ramesh A Gopinath. Gaussianization. In Advances in Neural Information Processing Systems, 2000.

[12] Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. A recurrent latent variable model for sequential data. In Advances in neural information processing systems, pages 2962–2970, 2015.

[13] Peter Dayan, Geoffrey E Hinton, Radford M Neal, and Richard S Zemel. The helmholtz machine. Neural computation, 7(5):889–904, 1995.

[14] Gustavo Deco and Wilfried Brauer. Higher order statistical decorrelation without information loss. In G. Tesauro, D. S. Touretzky, and T. K. Leen, editors, Advances in Neural Information Processing Systems 7, pages 247–254. MIT Press, 1995.

[15] Emily L. Denton, Soumith Chintala, Arthur Szlam, and Rob Fergus. Deep generative image models using a laplacian pyramid of adversarial networks. In Advances in Neural Information Processing Systems 28:

Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada, pages 1486–1494, 2015.

[16] Luc Devroye. Sample-based non-uniform random variate generation. In Proceedings of the 18th conference on Winter simulation, pages 260–265. ACM, 1986.

[17] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: non-linear independent components estimation. arXiv preprint arXiv:1410.8516, 2014.

[18] Brendan J Frey. Graphical models for machine learning and digital communication. MIT press, 1998.

[19] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Texture synthesis using convolutional neural networks. In Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada, pages 262–270, 2015.

[20] Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. MADE: masked autoencoder for distribution estimation. CoRR, abs/1502.03509, 2015.

[21] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada, pages 2672–2680, 2014.

[22] Karol Gregor, Frederic Besse, Danilo Jimenez Rezende, Ivo Danihelka, and Daan Wierstra. Towards conceptual compression. arXiv preprint arXiv:1604.08772, 2016.

[23] Shixiang Gu, Timothy Lillicrap, Ilya Sutskever, and Sergey Levine. Continuous deep q-learning with model-based acceleration. arXiv preprint arXiv:1603.00748, 2016.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. CoRR, abs/1512.03385, 2015.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. CoRR, abs/1603.05027, 2016.

[26] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. Neural Computation, 9(8):1735–1780, 1997.

[27] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. The Journal of Machine Learning Research, 14(1):1303–1347, 2013.

[28] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. Independent component analysis, volume 46. John Wiley & Sons, 2004.

[29] Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. Neural Networks, 12(3):429–439, 1999.

[30] Daniel Jiwoong Im, Chris Dongjoo Kim, Hui Jiang, and Roland Memisevic. Generating images with recurrent adversarial networks. arXiv preprint arXiv:1602.05110, 2016.

[31] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167, 2015.

[32] Rafal Józefowicz, Oriol Vinyals, Mike Schuster, Noam Shazeer, and Yonghui Wu. Exploring the limits of language modeling. CoRR, abs/1602.02410, 2016.

[33] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

[34] Diederik P Kingma, Tim Salimans, and Max Welling. Improving variational inference with inverse autoregressive flow. arXiv preprint arXiv:1606.04934, 2016.

[35] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013.

[36] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images, 2009.

[37] Hugo Larochelle and Iain Murray. The neural autoregressive distribution estimator. In AISTATS, 2011.

[38] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric. CoRR, abs/1512.09300, 2015.

[39] Yann A LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. Efficient backprop. In Neural networks: Tricks of the trade, pages 9–48. Springer, 2012.

[40] Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. arXiv preprint arXiv:1409.5185, 2014.

[41] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In Proceedings of International Conference on Computer Vision (ICCV), December 2015.

[42] Lars Maaløe, Casper Kaae Sønderby, Søren Kaae Sønderby, and Ole Winther. Auxiliary deep generative models. arXiv preprint arXiv:1602.05473, 2016.

[43] Andriy Mnih and Karol Gregor. Neural variational inference and learning in belief networks. arXiv preprint arXiv:1402.0030, 2014.

[44] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. Nature, 518(7540):529–533, 2015.

[45] Radford M Neal and Geoffrey E Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. In Learning in graphical models, pages 355–368. Springer, 1998.

[46] Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. Pixel recurrent neural networks. arXiv preprint arXiv:1601.06759, 2016.
[47] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. CoRR, abs/1511.06434, 2015.
[48] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. arXiv preprint arXiv:1505.05770, 2015.
[49] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approxi-mate inference in deep generative models. arXiv preprint arXiv:1401.4082, 2014.
[50] Oren Rippel and Ryan Prescott Adams. High-dimensional probability estimation with deep density models. arXiv preprint arXiv:1302.5125, 2013.
[51] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. Cognitive modeling, 5(3):1, 1988.
[52] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. International Journal of Computer Vision, 115(3):211–252, 2015.
[53] Ruslan Salakhutdinov and Geoffrey E Hinton. Deep boltzmann machines. In International conference on artificial intelligence and statistics, pages 448–455, 2009.
[54] Tim Salimans and Diederik P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. arXiv preprint arXiv:1602.07868, 2016.
[55] Tim Salimans, Diederik P Kingma, and Max Welling. Markov chain monte carlo and variational inference: Bridging the gap. arXiv preprint arXiv:1410.6460, 2014.
[56] Lawrence K Saul, Tommi Jaakkola, and Michael I Jordan. Mean field theory for sigmoid belief networks. Journal of artificial intelligence research, 4(1):61–76, 1996.
[57] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recogni-tion. arXiv preprint arXiv:1409.1556, 2014.
[58] Paul Smolensky. Information processing in dynamical systems: Foundations of harmony theory. Technical report, DTIC Document, 1986.
[59] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015, pages 2256–2265, 2015.
[60] Sasha Targ, Diogo Almeida, and Kevin Lyman. Resnet in resnet: Generalizing residual architectures. CoRR, abs/1603.08029, 2016.
[61] Lucas Theis and Matthias Bethge. Generative image modeling using spatial lstms. In Advances in Neural Information Processing Systems, pages 1918–1926, 2015.
[62] Lucas Theis, Aäron Van Den Oord, and Matthias Bethge. A note on the evaluation of generative models. CoRR, abs/1511.01844, 2015.
[63] Dustin Tran, Rajesh Ranganath, and David M Blei. Variational gaussian process. arXiv preprint arXiv:1511.06499, 2015.
[64] Benigno Uria, Iain Murray, and Hugo Larochelle. Rnade: The real-valued neural autoregressive density-estimator. In Advances in Neural Information Processing Systems, pages 2175–2183, 2013.
[65] Hado van Hasselt, Arthur Guez, Matteo Hessel, and David Silver. Learning functions across many orders of magnitudes. arXiv preprint arXiv:1602.07714, 2016.
[66] Oriol Vinyals, Samy Bengio, and Manjunath Kudlur. Order matters: Sequence to sequence for sets. arXiv preprint arXiv:1511.06391, 2015.
[67] Manuel Watter, Jost Springenberg, Joschka Boedecker, and Martin Riedmiller. Embed to control: A locally linear latent dynamics model for control from raw images. In Advances in Neural Information Processing Systems, pages 2728–2736, 2015.
[68] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine learning, 8(3-4):229–256, 1992.
[69] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122, 2015.
[70] Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint arXiv:1506.03365, 2015.
[71] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. arXiv preprint arXiv:1603.08511, 2016.

A 样本

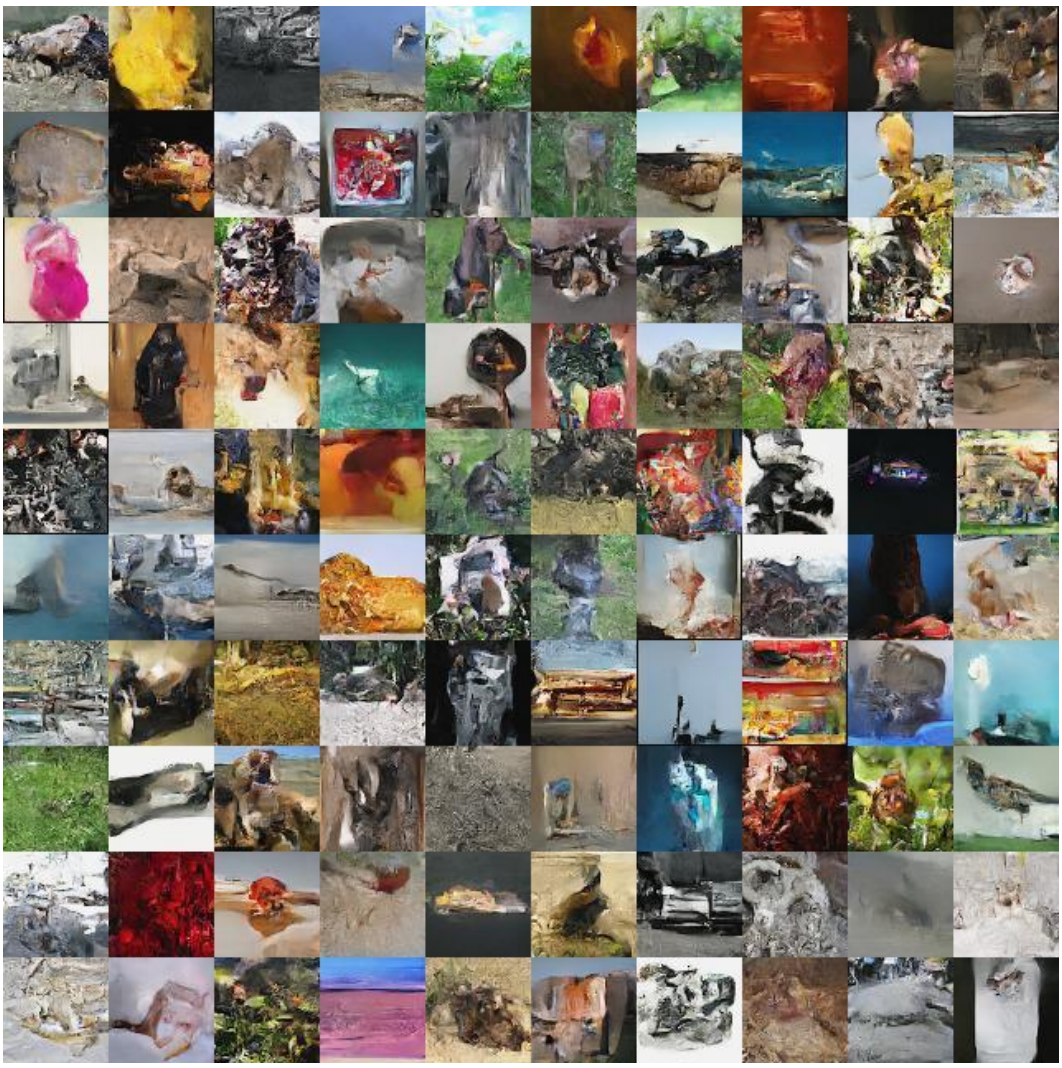


图 7：来自 Imagenet 训练模型的样本（64×64）。

图 8：来自 CelebA 训练模型的样本。

图 9：来自 LSUN（卧室类别）训练模型的样本。

图 10：来自 LSUN（教堂室外类别）训练模型的样本。

图 11：来自在 LSUN（塔类别）上训练的模型的样本。

B 流形

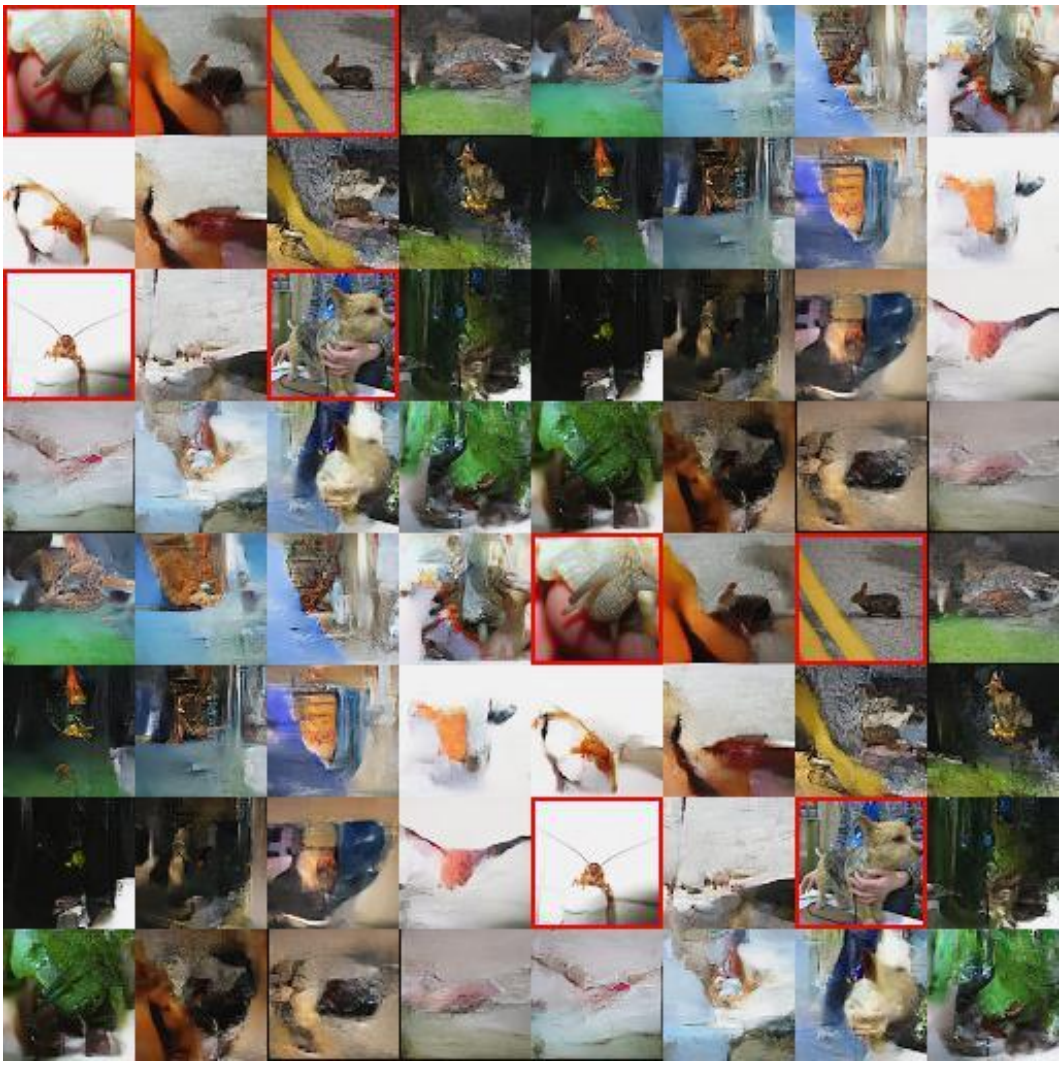


图 12：来自 Imagenet 训练模型的流形（64×64）。 带有红色边框的图像从验证集中获取，并定义流形。 如等式 19 中所述计算流形，其中 x 轴对应于φ，并且 y 轴对应于 ϕ'，并且其中$\phi, \phi' \in \{0, \frac{\pi}{4}, \cdots, \frac{7\pi}{4}\}$。

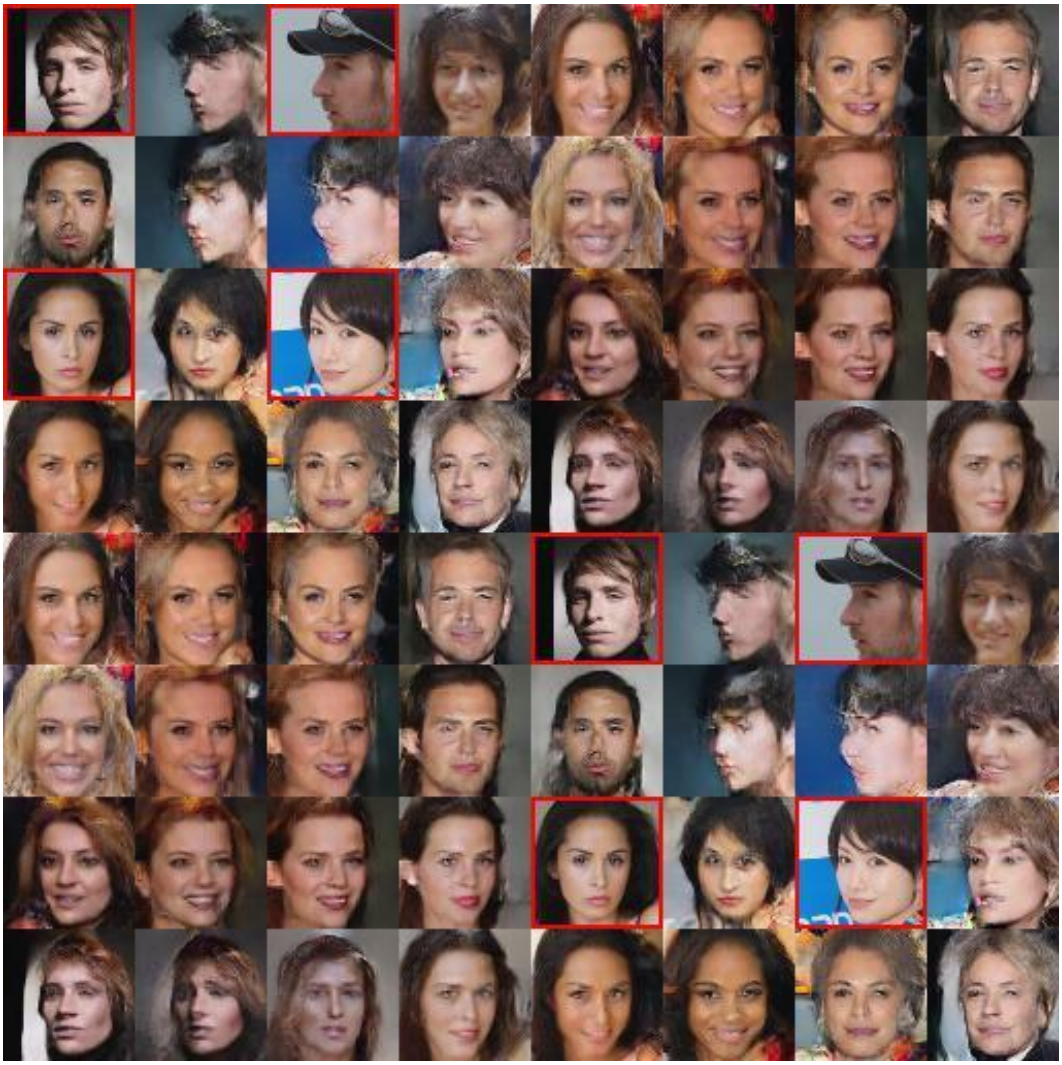


图 13：来自 CelebA 训练模型的流形。 带有红色边框的图像取自训练集，并定义流形。如等式 19 中所述计算流形，其中 x 轴对应于φ，并且 y 轴对应于 ϕ' ，并且其中$\phi, \phi' \in \{0, \frac{\pi}{4}, \cdots, \frac{7\pi}{4}\}$。

图 14：在 LSUN（卧室类别）上训练的模型的流形。 具有红色边界的图像从验证集中获取，并定义流形。 如等式 19 中所述计算流形，其中 x 轴对应于φ，并且 y 轴对应于 ϕ' ，并且其中$\phi, \phi' \in \{0, \frac{\pi}{4}, \cdots, \frac{7\pi}{4}\}$。

图 15：在 LSUN（教堂室外类别）训练的模型中的流形。 带有红色边框的图像从验证集中获取，并定义流形。如等式 19 中所述计算流形，其中 x 轴对应于φ，并且 y 轴对应于 ϕ'，并且其中$\phi, \phi' \in \{0, \frac{\pi}{4}, \cdots, \frac{7\pi}{4}\}$。

图 16：来自在 LSUN（塔类别）上训练的模型的流形。 具有红色边界的图像从验证集中获取，并定义流形。 如等式 19 中所述计算流形，其中 x 轴对应于φ，并且 y 轴对应于 ϕ'，并且其中$\phi, \phi' \in \{0, \frac{\pi}{4}, \cdots, \frac{7\pi}{4}\}$。

C 外推

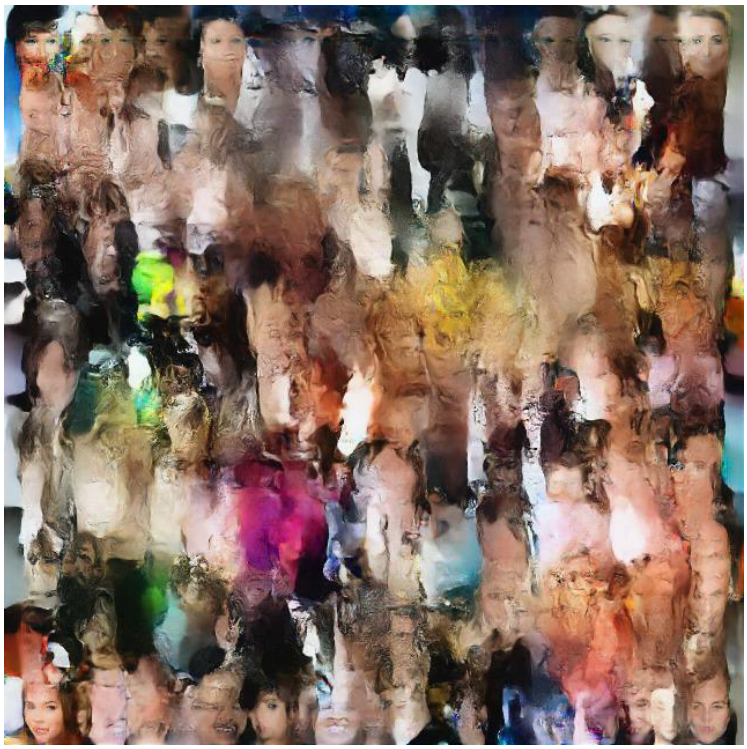
受[19,61]的纹理生成工作和 DCGAN 外推测试[47]的启发，我们还通过生成数据集中存在的图像的两倍或十倍来评估我们的模型捕获的统计数据。 正如我们在下图中所观察到的，我们的模型似乎成功地创建了数据集的“纹理”表示，同时保持了图像的空间平滑度。 我们的卷积结构只能通过卷积中的边缘效应知道所考虑像素的位置，因此我们的模型类似于静止过程。 这也解释了为什么这些样本在 LSUN 中更加一致，其中训练数据是使用随机作物获得的。

(a) 2

(b) 10

图 17：我们在 Imagenet 上生成比训练集图像大小更大的样本（64×64）。

(a) 2



(b) 10

图 18：我们在 CelebA 上生成的样本大于训练集图像大小。

(a) 2

(b) 10

图 19：我们在 LSUN（卧室类别）上生成比训练集图像大小更大的样本。

(a) 2

(b) 10

图 20：我们在 LSUN（教堂室外类别）上生成比训练集图像大小更大的样本。

(a) 2

(b) 10

图 21：我们在 LSUN（塔类别）上生成比训练集图像大小更大的样本。

D 潜在变量语义

如[22]所述，我们进一步尝试通过消融测试来掌握学习层潜在变量的语义。 我们推断潜在变量并从标准高斯重新采样潜在变量的最低水平，增加受此重采样影响的最高水平。 正如我们在下面的图中所看到的，我们潜在空间的语义似乎更多地是在图形层面而不是更高层次的概念。 虽然卷积的大量使用通过利用图像先验知识来改善学习，但也可能是造成这种限制的原因。

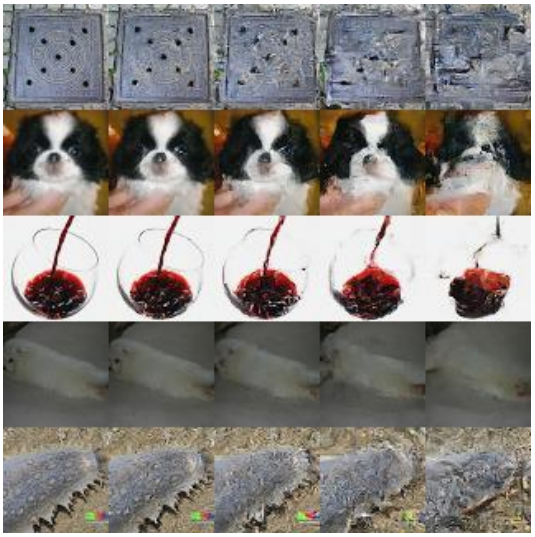


图 22：在 Imagenet 上训练的模型的概念压缩（64×64）。 最左边的列代表原始图像，随后的列是通过存储更高级别的潜在变量并重新采样其他列来获得的，随着我们的正确存储越来越少。 从左到右：保留潜在变量的 100%，50%，25%，12.5%和 6.25%。

图 23：来自 CelebA 训练模型的概念压缩。 最左边的列代表原始图像，随后的列是通过存储更高级别的潜在变量并重新采样其他列来获得的，随着我们的正确存储越来越少。 从左到右：保留潜在变量的 100%，50%，25%，12.5%和 6.25%。

图 24：来自在 LSUN（卧室类别）上训练的模型的概念压缩。 最左边的列代表原始图像，随后的列是通过存储更高级别的潜在变量并重新采样其他列来获得的，随着我们的正确存储越来越少。 从左到右：保留潜在变量的 100%，50%，25%，12.5%和 6.25%。

图 25：在 LSUN（教堂室外类别）上训练的模型的概念压缩。 最左边的列代表原始图像，随后的列是通过存储更高级别的潜在变量并重新采样其他列来获得的，随着我们的正确存储越来越少。 从左到右：保留潜在变量的 100%，50%，25%，12.5%和 6.25%。

图 26：来自在 LSUN（塔类别）上训练的模型的概念压缩。 最左边的列代表原始图像，随后的列是通过存储更高级别的潜在变量并重新采样其他列来获得的，随着我们的正确存储越来越少。 从左到右：保留潜在变量的 100%，50%，25%，12.5%和 6.25%。

E 批量归一化

我们通过使用层统计量的移动平均值的加权平均值来进一步试验批量归一化$\tilde{\mu}_t, \tilde{\sigma}_t^2$和当前批量统计$\hat{\mu}_t, \hat{\sigma}_t^2$,

$$\tilde{\mu}_{t+1} = \rho\tilde{\mu}_t + (1-\rho)\hat{\mu}_t \tag{20}$$

$$\tilde{\sigma}_{t+1}^2 = \rho\tilde{\sigma}_t^2 + (1-\rho)\hat{\sigma}_t^2, \tag{21}$$

其中 ρ 是动量。 使用$\tilde{\mu}_{t+1}, \tilde{\sigma}_{t+1}^2$时，我们只通过当前批量统计传播梯度$\hat{\mu}_t, \hat{\sigma}_t^2$。 我们观察到使用这种滞后有助于模型训练具有非常小的 minibatches。

我们在 CIFAR-10 上使用了具有移动平均值的批量标准化。

F 属性改变

另外，我们利用 CelebA 中的属性信息 y 来构建条件模型，即从图像到潜变量的可逆函数 f 使用 y 中的标签来定义其参数。 为了观察存在于潜在变量中的信息，我们选择使用其原始属性 y 对一批图像 x 进行编码，并使用一组新的属性 y0 对它们进行解码，通过对批次内的原始属性进行混洗来构建。 我们得到新的图像$x' = g\big(f(x;y);y'\big)$。

我们观察到，虽然面部被改变以贴合新属性，但是几个属性保持不变，如位置和背景。

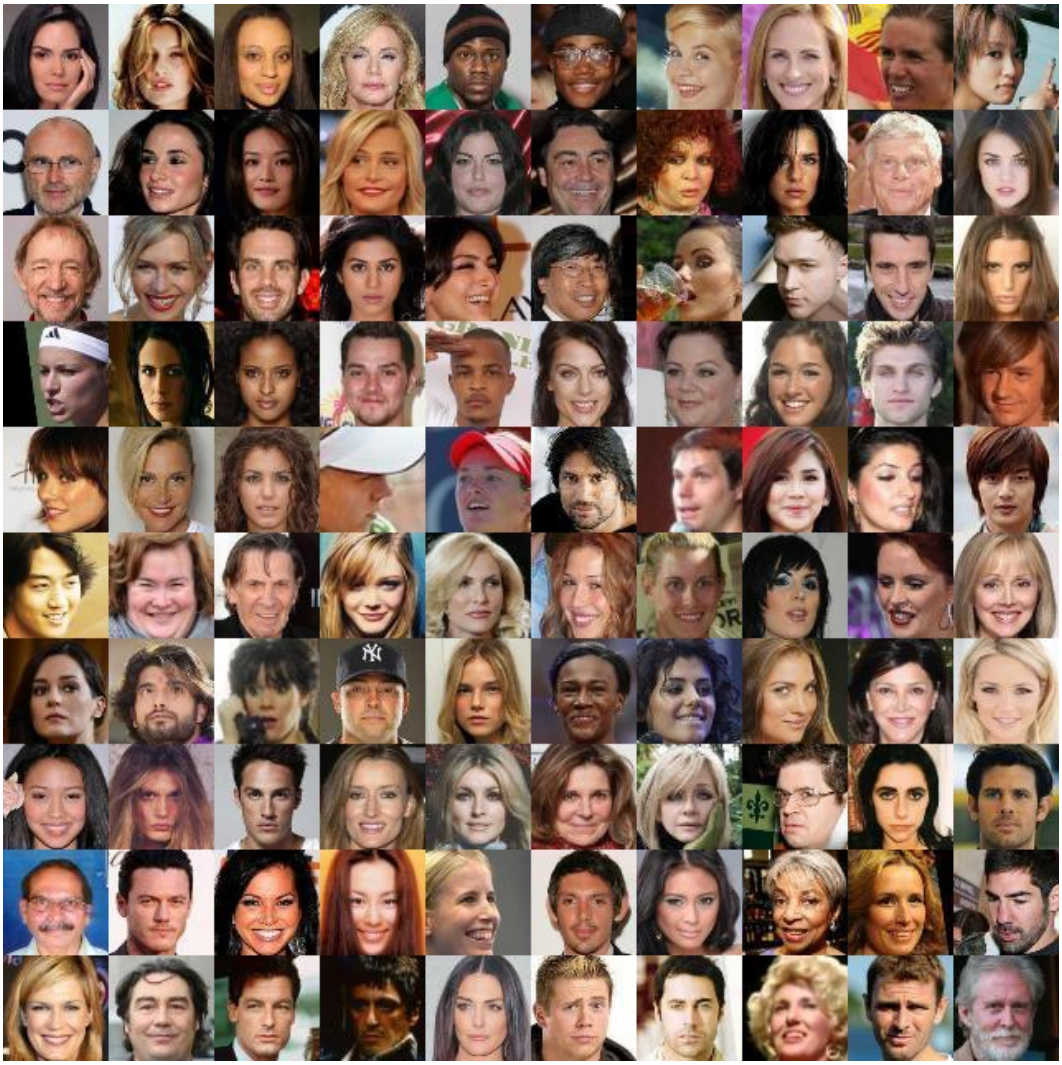


图 27：CelebA 数据集中的示例 x。

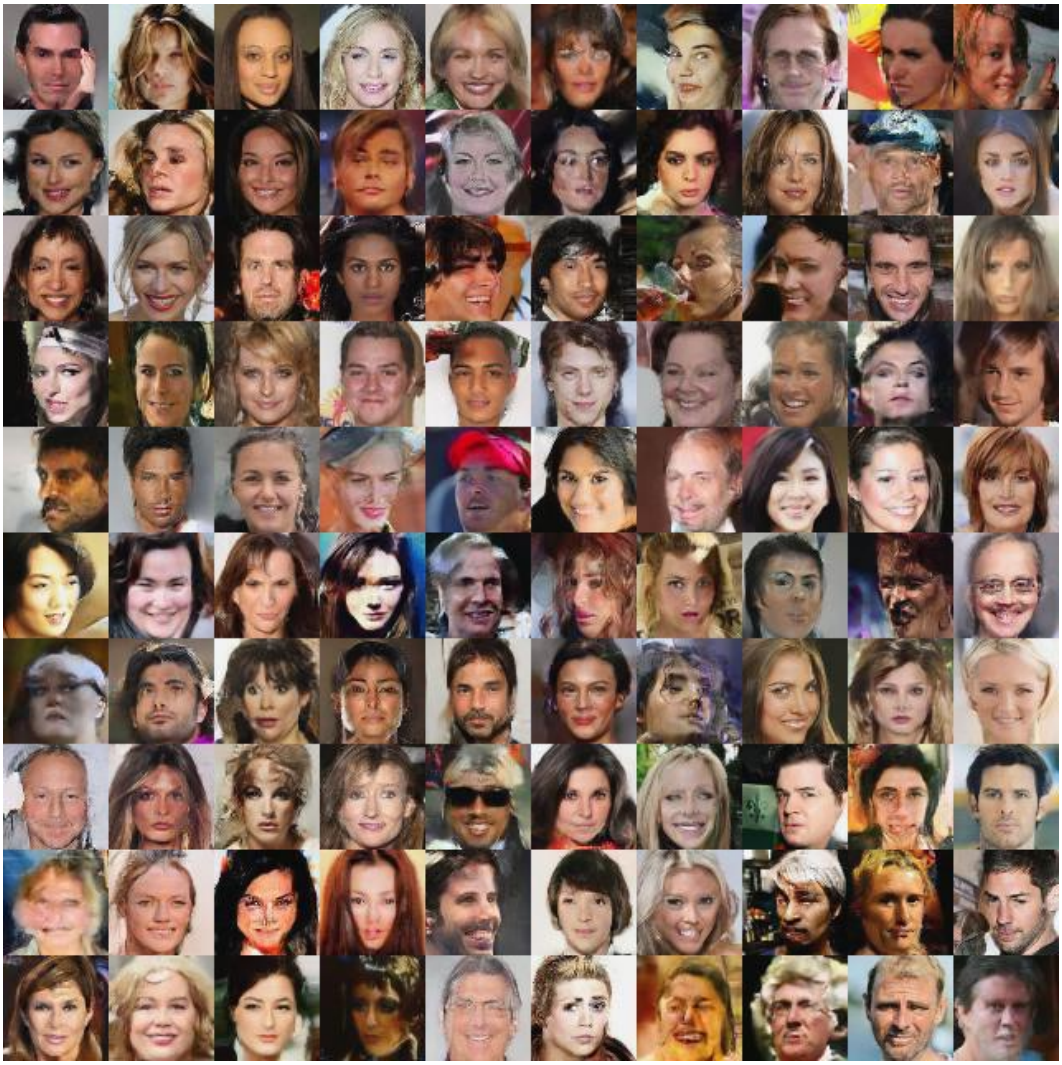


图 28：从使用 CelebA 数据集的图像和属性训练的模型中，我们使用一组新属性对一组图像进行原始属性编码。我们注意到新图像通常与图 27 中的图像具有相似的特征，包括位置和背景。