生成对抗网络

Ian J. Goodfellow, Jean Pouget-Abadie,[*] Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair,[†] Aaron Courville, Yoshua Bengio[‡]
Département d'informatique et de recherche opérationnelle
Université de Montréal
Montréal, QC H3C 3J7

Abstract

我们提出了一个估算生成模型的新框架通过对抗的过程，我们同时训练两个模型:一个生成模型G捕获数据分布，以及一个判别模型D来估计样本来自训练数据而不是G的概率。G的训练程序是使D的概率最大化一个错误。这个框架对应于一个极大极小的二人博弈。在空间任意函数G和D，存在唯一解，带G 恢复训练数据分布和D等于$\frac{1}{2}$ 到处都是。在G和D由多层感知器定义的情况下，整个系统可以通过反向传播进行训练。不需要任何马尔可夫链或者在训练或生成过程中展开近似推理网络样本。实验证明了该框架的潜力通过定性和定量的评价产生样品。

1 简介

深度学习的前景是发现丰富的分层模型[2]，这些模型表示在人工智能应用中遇到的各种数据的概率分布，例如自然图像，包含语音的音频波形，以及自然语言语料库中的符号。到目前为止，深度学习中最显著的成功都涉及到判别模型，通常是那些将高维、丰富的感官输入映射到班级的人标签[14, 22]。这些惊人的成功主要是基于反向传播和辍学算法，使用分段线性单位[19, 9, 10]有一个特别好的梯度．由于难以近似，深度生成的模型的影响较小许多难以处理的概率计算出现在最大似然估计和相关的策略，并且由于难以利用分段线性单元的好处生成环境。我们提出了一种新的生成模型估计方法来避免这些困难。[1]

在所提出的对抗网络框架中，生成模型与一个对手:一种判别模型，学习确定样本是否从模型分布或数据分布。可以考虑生成模型就像一群造假者，试图制造假币并使用它侦查，而判别模型则类似于警察，试图侦查假币。这个游戏中的竞争促使两队改进他们的方法，直到仿冒品出现与真品难以区分。

该框架可以为多种模型和优化生成特定的训练算法算法。在本文中，我们将探讨生成模型生成时的特殊情况样本通过多层感知器传递随机噪声，并采用判别模型也是一个多层感知器。我们把这种特殊情况称为对抗性网络。在这种情况下，我们可以只使用非常成功的反向传播和退出算法[17] 从生成式中抽样仅使用前向传播建模。不需要近似推理或马尔可夫链。

2 相关工作

具有潜在变量的有向图形模型的另一种选择是无向图形模型有潜在变量，比如受限玻尔兹曼机(rbm) [27, 16]，深度玻尔兹曼机(DBMs) [26]及其众多变体。这些模型中的交互被表示为非归一化的势函数，由一个全局归一化对随机变量的所有状态求和/积分。这个量(配

[*] Jean Pouget-Abadie正在访问巴黎综合理工大学。
[†] 来自德里印度理工学院的Sherjil Ozair正在访问蒙特里萨大学
[‡] Yoshua Bengio是CIFAR高级研究员。
[1] 所有代码和超参数可在`http://www.github.com/goodfeli/adversarial`

分函数)和除了最微不足道的，它的梯度对所有人来说都是棘手的实例，尽管它们可以用马尔可夫链Monte估计卡罗(MCMC)方法。混合给学习带来了很大的问题算法依赖于MCMC [3, 5]。

深度信念网络(dbn) [16]是包含单个无向层和多个有向层的混合模型。虽然存在一种快速的近似分层训练标准，但dbn会引起计算与无向和有向模型相关的困难。

不近似或限定的可选标准对数似然也被提出，如得分匹配[18] 和噪声对比估计(NCE) [13]。这两种方法都要求学习到的概率密度是解析指定的直到一个归一化常数。注意，在许多有趣的生成模型中有几个潜在变量层(如dbn和DBMs)，甚至不可能推导出可处理的非规范化概率密度。一些模型，如去噪自动编码器[30]和收缩自动编码器的学习规则与应用于rbm的分数匹配非常相似。在NCE中，就像在这项工作中一样，一个判别性训练标准是用来拟合生成模型。然而，与其单独拟合，不如区别对待模型，生成模型本身用于区分生成的数据和样本固定的噪声分布。因为NCE使用固定的噪声分布，所以学习变慢了在模型学习了一个近似正确的分布之后观察变量的一个小子集。

最后，一些技术不涉及明确定义概率分布，而是训练一个生成机器从期望的分布中抽取样本。这种方法的优点是，这样的机器可以被设计成由人来训练反向传播。最近在这一领域的突出工作包括生成随机网络(GSN)框架[5]，扩展了广义去噪自动编码器[4]:两者都可以被视为定义一个参数化马尔可夫链，即学习机器的参数执行生成马尔可夫链的一步。与GSNs相比，对抗网络框架不需要马尔可夫链抽样。因为对抗性网络在生成过程中不需要反馈回路，他们能够更好地利用分段线性单位[19, 9, 10]，它提高了反向传播的性能，但在反馈回路中使用时存在无界激活问题。最近的例子是通过反向传播来训练生成机器包括最近对自动编码变分贝叶斯的研究[20] 随机反向传播[24]。

3 对抗性网

当模型同时存在时，对抗性建模框架最容易应用多层感知器。为了了解生成器在数据$\boldsymbol{x}$上的分布p_g，我们定义输入噪声变量$p_{\boldsymbol{z}}(\boldsymbol{z})$的先验，然后表示a 映射到数据空间为$G(\boldsymbol{z};\theta_g)$，其中$G$是一个可微函数由参数为$\theta_g$的多层感知器表示。我们还定义了秒输出单个标量的多层感知器$D(\boldsymbol{x};\theta_d)$。$D(\boldsymbol{x})$表示$\boldsymbol{x}$来自数据而不是$p_g$的概率。我们训练$D$来最大化将正确标签分配给训练样例和来自$G$的样本的概率。我们同时训练$G$最小化$\log(1 - D(G(\boldsymbol{z})))$:

换句话说，D和G进行如下的二人极大极小博弈，其值函数为$V(G, D)$:

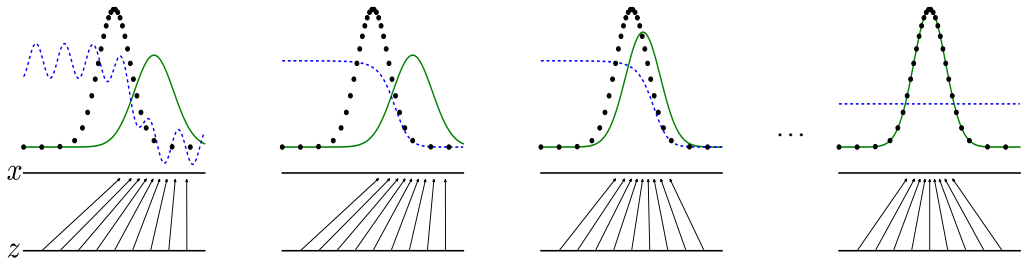
$$\min_G \max_D V(D, G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))]. \tag{1}$$

在下一节中，我们将对对抗网络进行理论分析，本质上表明训练标准允许恢复数据生成分布如G和D被赋予足够的容量，即在非参数极限。请参见图1，以获得不太正式，更具有教学意义的内容方法解释。在实践中，我们必须使用迭代的数值方法来执行游戏。优化D在完成训练的内循环在计算上是令人望而却步的，在有限的数据集上会导致过拟合。相反，我们在k步骤之间交替进行优化D和一步优化G。这导致维护D 接近最优解，只要G变化足够慢。这个策略是类似的到SML/PCD [31, 29]训练从一个马尔可夫链中维护样本的方式学习下一步，以避免在马尔可夫链中燃烧，作为内环的一部分关于学习。程序已正式提出在算法1。

在实践中，方程1可能不能为G提供足够的梯度嗯。在学习的早期，当G很差时，D可以拒绝高置信度的样本，因为它们是明显不同于训练数据。在这种情况下，$\log(1 - D(G(\boldsymbol{z})))$饱和。而不是训练$G$最小化$\log(1 - D(G(\boldsymbol{z})))$我们可以训练$G$最大化$\log D(G(\boldsymbol{z}))$。该目标函数的结果是$G$和$D$的动力学不动点相同，但提供了很多在学习早期有更强的梯度。

4 理论结果

生成器G隐式地定义一个概率分布p_g为得到的样本分布$G(\boldsymbol{z})\boldsymbol{z} \sim p_{\boldsymbol{z}}$。因此，我们希望算法1收敛于a 如果有足够的能力和训练时间，可以很好地估计p_{data}。The 这部分的结果是在非参数设置下完成的，例如我们表示a 通过研究概率空间中的收敛性，建立了具有无限容量的模型密度函数。



(a) (b) (c) (d)

Figure 1: 生成式对抗网络通过同步更新鉴别的 分布(D(蓝色，虚线)，以便区分来自的样本数据生成分布(黑色，虚线) $p_{\boldsymbol{x}}$ 从那些生成的 分布p_g (G)(绿色实线)。下面的水平线是$\boldsymbol{z}$ 是均匀抽样的。上面的水平线是的定义域的一部分$\boldsymbol{x}$。向上的箭头表示如何映射$\boldsymbol{x} = G(\boldsymbol{z})$ 施加非均匀分布p_g 在变换后的样本上。G 在高密度区域收缩，在低密度区域膨胀p_g。(a)考虑一个接近收敛的对抗性对: p_g 类似于p_{data} 和D 是一个部分准确的分类器。(b)在算法的内循环中D 被训练来区分样本和数据，收敛于$D^*(\boldsymbol{x}) = \frac{p_{\text{data}}(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x})+p_g(\boldsymbol{x})}$。(c)更新后$G$，梯度$D$ 引导了$G(\boldsymbol{z})$ 流向更有可能归类为数据(d)经过几个训练步骤后，如果G 和D 有了足够的能力，他们会达到什么地步两者都无法改善，因为$p_g = p_{\text{data}}$。鉴别器无法区分这两者分布，即$D(\boldsymbol{x}) = \frac{1}{2}$。

Algorithm 1 生成对抗网络的小批量随机梯度下降训练。应用到鉴别器k的步数是一个超参数。我们使用$k = 1$ 在我们的实验中，最便宜的选择。

for number of training iterations **do**

for k steps **do**

- Sample minibatch of m noise samples $\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(m)}\}$ from noise prior $p_g(\boldsymbol{z})$.
- Sample minibatch of m examples $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$ from data generating distribution $p_{\text{data}}(\boldsymbol{x})$.
- Update the discriminator by ascending its stochastic gradient:

$$\nabla_{\theta_d} \frac{1}{m} \sum_{i=1}^{m} \left[\log D\left(\boldsymbol{x}^{(i)}\right) + \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right) \right].$$

end for

- Sample minibatch of m noise samples $\{\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(m)}\}$ from noise prior $p_g(\boldsymbol{z})$.
- Update the generator by descending its stochastic gradient:

$$\nabla_{\theta_g} \frac{1}{m} \sum_{i=1}^{m} \log\left(1 - D\left(G\left(\boldsymbol{z}^{(i)}\right)\right)\right).$$

end for

The gradient-based updates can use any standard gradient-based learning rule. We used momentum in our experiments.

我们将在4.1节中展示这个极大极小博弈对$p_g = p_{\text{data}}$有全局最优。然后我们就进来章节4.2该算法1 对Eq 1进行优化，从而得到所期望的结果。

4.1 的全局最优性$p_g = p_{\text{data}}$

我们首先考虑任意给定生成器G的最优鉴别器D。

Proposition 1. *For G fixed, the optimal discriminator D is*

$$D_G^*(\boldsymbol{x}) = \frac{p_{data}(\boldsymbol{x})}{p_{data}(\boldsymbol{x}) + p_g(\boldsymbol{x})} \tag{2}$$

Proof. 对于任意生成器G，鉴别器D的训练准则是使数量最大化$V(G,D)$

$$\begin{aligned} V(G,D) &= \int_{\boldsymbol{x}} p_{\text{data}}(\boldsymbol{x}) \log(D(\boldsymbol{x}))dx + \int_{z} p_{\boldsymbol{z}}(\boldsymbol{z}) \log(1-D(g(\boldsymbol{z})))dz \\ &= \int_{\boldsymbol{x}} p_{\text{data}}(\boldsymbol{x}) \log(D(\boldsymbol{x})) + p_g(\boldsymbol{x}) \log(1-D(\boldsymbol{x}))dx \end{aligned} \tag{3}$$

对于任意$(a,b) \in \mathbb{R}^2 \setminus \{0,0\}$，函数$y \to a\log(y) + b\log(1-y)$在$[0,1]$处达到最大值，为$\frac{a}{a+b}$。鉴别符不需要在$Supp(p_{\text{data}}) \cup Supp(p_g)$之外定义，从而得出证明。 □

请注意，D的训练目标可以解释为最大化用于估计条件概率$P(Y=y|\boldsymbol{x})$的对数似然，其中Y表示$\boldsymbol{x}$是来自p_{data}(含$y=1$)还是p_g(含$y=0$)。在Eq. 1中的极大极小游戏现在可以被重新表述为:

$$\begin{aligned} C(G) =& \max_D V(G,D) \\ =& \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}[\log D_G^*(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}}[\log(1-D_G^*(G(\boldsymbol{z})))] \\ =& \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}[\log D_G^*(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim p_g}[\log(1-D_G^*(\boldsymbol{x}))] \\ =& \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}\left[\log \frac{p_{\text{data}}(\boldsymbol{x})}{P_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}\right] + \mathbb{E}_{\boldsymbol{x}\sim p_g}\left[\log \frac{p_g(\boldsymbol{x})}{p_{\text{data}}(\boldsymbol{x}) + p_g(\boldsymbol{x})}\right] \end{aligned} \tag{4}$$

Theorem 1. *The global minimum of the virtual training criterion $C(G)$ is achieved if and only if $p_g = p_{data}$. At that point, $C(G)$ achieves the value $-\log 4$.*

Proof. 对于$p_g = p_{\text{data}}$, $D_G^*(\boldsymbol{x}) = \frac{1}{2}$，(考虑等式2)。因此，通过检查$D_G^*(\boldsymbol{x}) = \frac{1}{2}$上的等式4，我们找到了$C(G) = \log\frac{1}{2} + \log\frac{1}{2} = -\log 4$。要看到这是$C(G)$的最佳可能值，只有$p_g = p_{\text{data}}$才能达到，请观察

$$\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}[-\log 2] + \mathbb{E}_{\boldsymbol{x}\sim p_g}[-\log 2] = -\log 4$$

通过$C(G) = V(D_G^*, G)$减去这个表达式，我们得到:

$$C(G) = -\log(4) + KL\left(p_{\text{data}} \middle\| \frac{p_{\text{data}} + p_g}{2}\right) + KL\left(p_g \middle\| \frac{p_{\text{data}} + p_g}{2}\right) \tag{5}$$

其中KL是Kullback- Leibler散度。在前面的表达式中，我们认识到模型分布和数据生成过程之间存在Jensen—Shannon分歧:

$$C(G) = -\log(4) + 2 \cdot JSD\left(p_{\text{data}} \| p_g\right) \tag{6}$$

因为两个分布之间的Jensen- Shannon散度总是非负的，只有零当它们相等时，我们已经证明$C^* = -\log(4)$是$C(G)$和的全局最小值唯一的解决方案是$p_g = p_{\text{data}}$，即生成模型完美地复制数据生成过程。 □

4.2 算法收敛性1

Proposition 2. *If G and D have enough capacity, and at each step of Algorithm 1, the discriminator is allowed to reach its optimum given G, and p_g is updated so as to improve the criterion*

$$\mathbb{E}_{\boldsymbol{x}\sim p_{data}}[\log D_G^*(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{x}\sim p_g}[\log(1-D_G^*(\boldsymbol{x}))]$$

then p_g converges to p_{data}

Proof. 考虑$V(G,D) = U(p_g, D)$作为p_g的函数，如上述标准所做的那样。注意$U(p_g, D)$在p_g中是凸的。凸函数的上零点的子导数包括函数在最大值处的导数。换句

Model	MNIST	TFD
DBN [3]	138 ± 2	1909 ± 66
Stacked CAE [3]	121 ± 1.6	$\mathbf{2110 \pm 50}$
Deep GSN [6]	214 ± 1.1	1890 ± 29
Adversarial nets	$\mathbf{225 \pm 2}$	$\mathbf{2057 \pm 26}$

Table 1: 基于窗口的对数似然估计。MNIST上报告的数字是测试集上样本的平均对数似然，用样本均值的标准误差计算。在TFD上，我们计算了数据集折叠间的标准误差，使用验证集选择了不同的σ 每个折叠。在TFD上，对每个折叠进行σ交叉验证，并计算每个折叠的平均对数似然。对于MNIST，我们与数据集的实值(而不是二进制)版本的其他模型进行比较。

话说，如果$f(x) = \sup_{\alpha \in \mathcal{A}} f_\alpha(x)$和$f_\alpha(x)$在$x$中对于每个$\alpha$都是凸的，那么$\partial f_\beta(x) \in \partial f$如果$\beta = \arg\sup_{\alpha \in \mathcal{A}} f_\alpha(x)$。这相当于在给定相应的$G$的最优$D$处计算$p_g$的梯度下降更新。$\sup_D U(p_g, D)$在$p_g$中是凸的，并且在Thm 1中证明了一个唯一的全局最优，因此，对于p_g进行足够小的更新，p_g收敛到p_x，从而结束了证明。 □

在实践中，对抗性网络通过函数$G(\boldsymbol{z}; \theta_g)$表示有限的$p_g$分布族，我们优化$\theta_g$而不是$p_g$本身。使用多层感知器定义$G$引入了多个参数空间中的临界点。然而，多层的优异性能感知器在实践中表明，尽管它们缺乏理论保证，但它们是一种合理的模型。

5 实验

我们用一系列数据集训练了对抗网络，包括MNIST [23] 多伦多人脸数据库(TFD) [28]和CIFAR-10 [21]。发电机网采用混合整流线性激活[19, 9]和sigmoid 激活，而鉴别器网络使用maxout [10]激活。Dropout [17]应用于培训鉴别网。而我们的理论框架允许使用dropout和其他噪声在发生器的中间层，我们只使用噪声作为最底层的输入发电机网络的。

我们通过对样本拟合高斯Parzen窗口来估计p_g下测试集数据的概率使用G生成并报告该分布下的对数似然。高斯函数的σ参数在验证集上进行交叉验证得到。这个过程在Breuleux *et al.* [8]中有介绍并用于各种生成模型，这些模型的确切似然是不易于处理[25, 3, 5]。结果见表1。这种估计可能性的方法有很大的方差在高维空间中表现不佳但这是我们已知的最好的方法。生成模型的进步，可以抽样，但不能估计可能性直接激励进一步的研究如何评估这些模型。

在图2和3中展示训练后从生成器网中抽取的样本。但我们并没有声称这些样品比用现有方法生成的样本，我们认为这些样本在与文献中更好的生成模型相比最不具竞争力强调对抗性框架的潜力。

6 优点和缺点

与以前的框架相比，这个新框架有优点也有缺点建模框架。缺点主要是没有明确的表示$p_g(\boldsymbol{x})$，并且D必须与G很好地同步在培训期间(特别是G)一定不要训练过多而不更新D，以避免"Helvetica场景"，其中G崩溃太多将$\mathbf{z}$的值与$\mathbf{x}$的值相同，以具有足够的多样性建模p_{data})，就像玻尔兹曼机的负链一样在学习步骤之间保持最新状态。它的优点是马尔可夫链是不需要的，只使用反向支撑来获得梯度，学习过程中不需要推理，而且种类繁多功能可以合并到模型中。表2总结了生成对抗网络与其他网络的比较生成建模方法。

上述优点主要是计算性的。对抗性模型可能会也获得了一些统计上的优势，从发电机网络不更新直接使用数据示例，但只能使用流过鉴别器的梯度。这意味着输入的组件不会直接复制到生成器的组件中参数。对抗性网络的另一个优势是它们可以代表非常尖锐，甚至退化的分布，而基于马尔可夫链的方法需要为了使链能够混合，它的分布有些模糊模式之间。

7 结论及未来工作

这个框架允许许多简单的扩展:

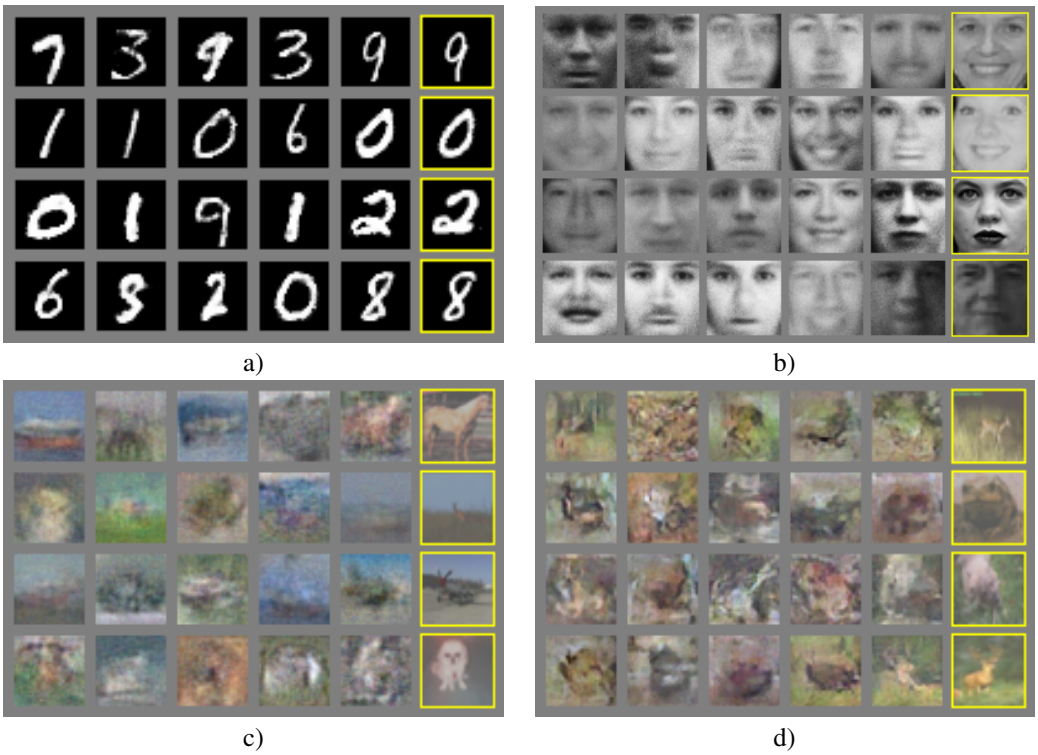


Figure 2: 模型中样本的可视化。最右边的列显示邻近样本的最近训练样例，以便演示模型没有记住训练集。样本是完全随机的抽签，不是精心挑选的。不像大多数其他的深度生成可视化模型，这些图像显示了模型分布的实际样本，而不是条件是指给定的隐藏单元样本。此外，这些样本是不相关，因为采样过程不依赖于马尔可夫链混合。a) MNIST b) TFD c) CIFAR-10(全连接模型) d) CIFAR-10(卷积鉴别器和"反卷积"生成器)

Figure 3: 之间线性插值得到的数字完整模型在$\boldsymbol{z}$空间中的坐标。

1. 通过添加$\boldsymbol{c}$，可以得到一个模型$p(\boldsymbol{x} \mid \boldsymbol{c})$ 作为G和D的输入。
2. 学习到的近似推理可以通过训练辅助网络来进行预测$\boldsymbol{z}$给定$\boldsymbol{x}$。这类似于通过唤醒-睡眠算法[15]训练的推理网络而推理网的优点是可以在生成后对固定的生成网进行训练Net已经完成了训练。
3. 可以近似地对所有条件$p(\boldsymbol{x}_S \mid \boldsymbol{x}_{\mathcal{S}})$建模，其中$S$是的子集$\boldsymbol{x}$的指数通过训练一组共享参数的条件模型。本质上，我们可以使用对抗网络来实现确定性MP-DBM的随机扩展[11]。
4. 半监督学习:来自鉴别器或推理网络的特征可以提高性能当可用的标记数据有限时，分类器的。
5. 效率的提高:通过划分更好的方法，培训可以大大加快协调G和D，或在培训期间确定更好的分布来采样$\mathbf{z}$。

本文论证了对抗性建模框架的可行性，建议这些研究路牌可能会很有用。

致谢

我们要感谢Patrice Marcotte、Olivier Delalleau、Kyunghyun Cho、Guillaume Alain和Jason Yosinski的有益讨论。Yann Dauphin与我们分享了他的Parzen窗口计算代码。我们要感谢的开发者Pylearn2 [12] 和特别是Theano [7, 1] fracimazric Bastien，他专门为这个项目开发了一个Theano特性。Arnaud Bergeron为LaTeX排版提供了急需的支持。我们还要感谢CIFAR和加拿大研究主席的资助，以及加拿大计算和计算quacimubec 用于提供计算资源。伊恩·古德

	Deep directed graphical models	Deep undirected graphical models	Generative autoencoders	Adversarial models
Training	Inference needed during training.	Inference needed during training. MCMC needed to approximate partition function gradient.	Enforced tradeoff between mixing and power of reconstruction generation	Synchronizing the discriminator with the generator. Helvetica.
Inference	Learned approximate inference	Variational inference	MCMC-based inference	Learned approximate inference
Sampling	No difficulties	Requires Markov chain	Requires Markov chain	No difficulties
Evaluating $p(x)$	Intractable, may be approximated with AIS	Intractable, may be approximated with AIS	Not explicitly represented, may be approximated with Parzen density estimation	Not explicitly represented, may be approximated with Parzen density estimation
Model design	Nearly all models incur extreme difficulty	Careful design needed to ensure multiple properties	Any differentiable function is theoretically permitted	Any differentiable function is theoretically permitted

Table 2: 生成建模中的挑战:总结了每种方法的深度生成建模的不同方法所遇到的困难主要操作涉及一个模型。

费洛(Ian Goodfellow)获得了2013年谷歌深度奖学金的支持学习。最后，我们要感谢三兄弟激发了我们的创造力。

References

[1] Bastien, F., Lamblin, P., Pascanu, R., Bergstra, J., Goodfellow, I. J., Bergeron, A., Bouchard, N., and Bengio, Y. (2012). Theano: new features and speed improvements. Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop.

[2] Bengio, Y. (2009). *Learning deep architectures for AI*. Now Publishers.

[3] Bengio, Y., Mesnil, G., Dauphin, Y., and Rifai, S. (2013a). Better mixing via deep representations. In *ICML'13*.

[4] Bengio, Y., Yao, L., Alain, G., and Vincent, P. (2013b). Generalized denoising auto-encoders as generative models. In *NIPS26*. Nips Foundation.

[5] Bengio, Y., Thibodeau-Laufer, E., and Yosinski, J. (2014a). Deep generative stochastic networks trainable by backprop. In *ICML'14*.

[6] Bengio, Y., Thibodeau-Laufer, E., Alain, G., and Yosinski, J. (2014b). Deep generative stochastic networks trainable by backprop. In *Proceedings of the 30th International Conference on Machine Learning (ICML'14)*.

[7] Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., Turian, J., Warde-Farley, D., and Bengio, Y. (2010). Theano: a CPU and GPU math expression compiler. In *Proceedings of the Python for Scientific Computing Conference (SciPy)*. Oral Presentation.

[8] Breuleux, O., Bengio, Y., and Vincent, P. (2011). Quickly generating representative samples from an RBM-derived process. *Neural Computation*, **23**(8), 2053–2073.

[9] Glorot, X., Bordes, A., and Bengio, Y. (2011). Deep sparse rectifier neural networks. In *AISTATS'2011*.

[10] Goodfellow, I. J., Warde-Farley, D., Mirza, M., Courville, A., and Bengio, Y. (2013a). Maxout networks. In *ICML'2013*.

[11] Goodfellow, I. J., Mirza, M., Courville, A., and Bengio, Y. (2013b). Multi-prediction deep Boltzmann machines. In *NIPS'2013*.

[12] Goodfellow, I. J., Warde-Farley, D., Lamblin, P., Dumoulin, V., Mirza, M., Pascanu, R., Bergstra, J., Bastien, F., and Bengio, Y. (2013c). Pylearn2: a machine learning research library. *arXiv preprint arXiv:1308.4214*.

[13] Gutmann, M. and Hyvarinen, A. (2010). Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *AISTATS'2010*.

[14] Hinton, G., Deng, L., Dahl, G. E., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., and Kingsbury, B. (2012a). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, **29**(6), 82–97.

[15] Hinton, G. E., Dayan, P., Frey, B. J., and Neal, R. M. (1995). The wake-sleep algorithm for unsupervised neural networks. *Science*, **268**, 1558–1161.

[16] Hinton, G. E., Osindero, S., and Teh, Y. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, **18**, 1527–1554.

[17] Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2012b). Improving neural networks by preventing co-adaptation of feature detectors. Technical report, arXiv:1207.0580.

[18] Hyvärinen, A. (2005). Estimation of non-normalized statistical models using score matching. *J. Machine Learning Res.*, **6**.

[19] Jarrett, K., Kavukcuoglu, K., Ranzato, M., and LeCun, Y. (2009). What is the best multi-stage architecture for object recognition? In *Proc. International Conference on Computer Vision (ICCV'09)*, pages 2146–2153. IEEE.

[20] Kingma, D. P. and Welling, M. (2014). Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[21] Krizhevsky, A. and Hinton, G. (2009). Learning multiple layers of features from tiny images. Technical report, University of Toronto.

[22] Krizhevsky, A., Sutskever, I., and Hinton, G. (2012). ImageNet classification with deep convolutional neural networks. In *NIPS'2012*.

[23] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, **86**(11), 2278–2324.

[24] Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. Technical report, arXiv:1401.4082.

[25] Rifai, S., Bengio, Y., Dauphin, Y., and Vincent, P. (2012). A generative process for sampling contractive auto-encoders. In *ICML'12*.

[26] Salakhutdinov, R. and Hinton, G. E. (2009). Deep Boltzmann machines. In *AISTATS'2009*, pages 448–455.

[27] Smolensky, P. (1986). Information processing in dynamical systems: Foundations of harmony theory. In D. E. Rumelhart and J. L. McClelland, editors, *Parallel Distributed Processing*, volume 1, chapter 6, pages 194–281. MIT Press, Cambridge.

[28] Susskind, J., Anderson, A., and Hinton, G. E. (2010). The Toronto face dataset. Technical Report UTML TR 2010-001, U. Toronto.

[29] Tieleman, T. (2008). Training restricted Boltzmann machines using approximations to the likelihood gradient. In W. W. Cohen, A. McCallum, and S. T. Roweis, editors, *ICML 2008*, pages 1064–1071. ACM.

[30] Vincent, P., Larochelle, H., Bengio, Y., and Manzagol, P.-A. (2008). Extracting and composing robust features with denoising autoencoders. In *ICML 2008*.

[31] Younes, L. (1999). On the convergence of Markovian stochastic algorithms with rapidly decreasing ergodicity rates. *Stochastics and Stochastic Reports*, **65**(3), 177–228.