基于能量的生成对抗网络

Junbo Zhao, Michael Mathieu and Yann LeCun
Department of Computer Science, New York University
Facebook Artificial Intelligence Research
{jakezhao, mathieu, yann}@cs.nyu.edu

摘 要

我们引入了“基于能量的生成对抗网络”模型（EBGAN），该模型将鉴别器视为能量函数，其将低能量归因于数据流形附近的区域以及将更高能量归因于其他区域。与基于概率的 GAN 类似，生成器被视为训练以产生具有最小能量的对比样本，而鉴别器被训练为将高能量分配给这些生成的样本。将判别方法视为能量函数除了具有逻辑斯蒂输出的通常二元分类器之外，还允许使用各种体系结构和损失函数。其中，我们将 EBGAN 框架的一个实例显示为使用自动编码器架构，其中能量变成重建损失，代替了判别器。我们表明，这种形式的 EBGAN 在训练期间表现出比常规 GAN 更稳定的行为。我们还展示了可以训练单尺度架构来生成高分辨率图像。

1 介绍

1.1 基于能量的模型

基于能量的模型（LeCun 等，2006）的本质是构建一个函数，将输入空间的每个点映射到单个标量，称为“能量”。 学习阶段是一个数据驱动的过程，它以一种方式塑造能量表面，使得所需的配置得到低能量的分配，而不正确的配置得到高能量。 监督学习属于这个框架：对于训练集中的每个 X，当 Y 是正确的标签时，对应的能量（X,Y）取低值，而对于不正确的 Y 则取较高的值。 类似地，当在无监督学习设置内单独建模 X 时，较低的能量归因于数据流形。 术语“对比样本”通常用于指引起能量上拉的数据点，例如监督学习中的不正确 Y 和无监督学习中来自低数据密度区域的点。

1.2 生成对抗网络

生成性对抗网络（GAN）（Goodfellow 等人，2014）已经导致图像生成的显著改善（Denton 等人，2015; Radford 等人，2015; Im 等人，2016; Salimans 等人，2016），视频预测（Mathieu 等，2015）和许多其他领域。 GAN 的基本思想是同时训练鉴别器和生成器。训练鉴别器以区分数据集的真实样本与生成器产生的假样本。该生成器使用来自易于采样的随机源的输入，并且经过训练以产生鉴别器无法与实际数据样本区分的假样本。在训练期间，生成器接收鉴别器的输出相对于假样本的梯度。在 Goodfellow 等人的 GAN 原始配方中（2014），鉴别器产生概率，并且在某些条件下，当由生成器产生的分布与数据分布匹配时发生收敛。从博弈论的角度来看，当生成器和鉴别器达到纳什均衡时，就达到了 GAN 的收敛。

1.3 基于能量的生成对抗网络

在这项工作中，我们建议在没有明确的概率解释的情况下将鉴别器视为能量函数（或对比函数）。由鉴别器计算的能量函数可以被视为生成器的可训练代价函数。训练鉴别器以将低能量值分配给高数据密度的区域，以及在这些区域之外的较高能量值。相反，生成器可以被视为可训练的参数化函数，其在鉴别器分配低能量的空间区域中产生样本。虽然通常可以通过 Gibbs 分布将能量转换为概率（LeCun 等人，2006），但是这种基于能量的 GAN 形式的归一化不能为鉴别器的结构选择和训练程序提供更大的灵活性 。

如 LeCun 等人所述，GAN 的原始公式中的概率二元鉴别器可被视为定义对比度函数和损失函数的许多方法中的一种（2006 年）对于监督和弱监督学习而言，Ranzato 等（2007 年）对于无监督学习而言。我们通过实验证明了这一概念，在鉴别器是自动编码器架构的环境中，能量是重建误差。附录 B 中提供了 EBGAN 解释的更多细节。

我们的主要贡献总结如下：

- 用于生成对抗训练的基于能量的方法。
- 证明在简单的铰链损失下，当系统达到收敛时，EBGAN 的生成器产生遵循基础数据分布的点。
- 具有使用自动编码器架构的鉴别器的 EBGAN 框架，其中能量是重建误差。
- 一组系统实验，探索超参数和架构选择，为 EBGAN 和概率 GAN 产生良好的结果。
- 这是一个演示，EBGAN 框架可用于以 256 × 256 像素分辨率从 ImageNet 数据集生成具有合理外观的高分辨率图像，无需多尺度方法。

2 EBGAN 模型

设 p_{data} 是生成数据集的分布的基础概率密度。 训练生成器 G 以从随机矢量 z 产生样本 G（z），例如图像，该随机矢量 z 从已知分布 p_z 采样，例如 $\mathcal{N}(0,1)$。 鉴别器 D 采用实际或生成的图像，并相应地估计能量值 $E \in \mathbb{R}$，如后边所述。 为简单起见，我们假设 D 产生非负值，但只要值限制在下文中所述的，分析就会依然成立。

2.1 目标函数

鉴别器的输出经过目标函数以便形成能量函数，将低能量归因于实际数据样本并且将更高能量归因于生成的（“假的”）样本。 在这项工作中，我们使用了一种边际损失（margin loss），但正如 LeCun 等人所解释的那样，许多其他选择都是可能的（2006 年）。 与使用概率 GAN（Goodfellow 等，2014）所做的相似，我们使用两种不同的损失，一种用于训练 D 而另一种用于训练 G，以便在生成器远离收敛时获得更好质量的梯度。

给定正边际 m，数据样本 x 和生成样本 G（z），鉴别器损失 $\mathcal{L}_D$ 和生成器损失 $\mathcal{L}_G$ 由以下形式正式定义：

$$\mathcal{L}_D(x,z) = D(x) + [m - D\big(G(z)\big)]^+ \tag{1}$$

$$\mathcal{L}_G(z) = D\big(G(z)\big) \tag{2}$$

其中 $[\cdot]^+ = max(0,\cdot)$。 关于参数 G 的最小化 $\mathcal{L}_G$ 类似于最大化第二项 $\mathcal{L}_D$。 当 $D(G(z)) \geq m$ 时，它具有相同的最小但非零梯度。

2.2 解决方案的最优性

在本节中，我们将对 2.1 节中介绍的系统进行理论分析。 我们证明，如果系统达到纳什均衡，则生成器 G 产生的样本与数据集的分布不可分割。 该部分在非参数的设置中完成，即我们假设 D 和 G 具有无限容量。

给定生成器 G，令 p_G 为 G（z）的密度分布，其中 $z \sim p_z$。 换句话说，p_G 是由 G 生成的样本的密度分布。

我们定义 $V(G,D) = \int_{x,z} \mathcal{L}_D(x,z) p_{data}(x) p_z(z) \mathrm{d}x \mathrm{d}z$ 和 $U(G,D) = \int_z \mathcal{L}_G(z) p_z(z) \mathrm{d}z$。 我们训练鉴别器 D 以使量 V 最小化和训练生成器 G 以使量 U 最小化。系统的纳什均衡是一对 (G^*, D^*) 满足：

$$V(G^*, D^*) \leq V(G^*, D) \qquad \forall D \tag{3}$$

$$U(G^*, D^*) \leq U(G, D^*) \qquad \forall G \tag{4}$$

定理 1。如果 (D^*, G^*) 是系统的纳什均衡，则 $p_{G^*} = p_{data}$ 几乎无处不在，并且 $V(D^*, G^*) = m$。

证明。首先，我们观察到

$$V(G^*, D) = \int_x p_{data}(x) D(x) \mathrm{d}x + \int_z p_z(z) \left[m - D(G^*(z)) \right]^+ \mathrm{d}z \tag{5}$$

$$= \int_x \left(p_{data}(x) D(x) + p_{G^*}(x) \left[m - D(x) \right]^+ \right) \mathrm{d}x. \tag{6}$$

函数 $\varphi(y) = ay + b(m-y)^+$ 的分析（详见附录 A 中的引理 1）表明：

(a) $D^*(x) \leq m$ 几乎无处不在。 为了验证它，让我们假设存在一组非零的度量，使得 $D^*(x) > m$。令 $\tilde{D}(x) = \min(D^*(x), m)$，于是 $V(G^*, \tilde{D}) < V(G^*, D^*)$ 违反等式 3。

(b) 如果 a <b，则函数φ以 m 为最小值，否则φ的最小值为 0。 因此，当我们用这些值代替 $D^*(x)$ 时，$V(G^*, D)$ 达到最小值。 我们得到

$$V(G^*, D^*) = m \int_x \mathbb{1}_{p_{data}(x) < p_{G^*}(x)} p_{data}(x) \mathrm{d}x + m \int_x \mathbb{1}_{p_{data}(x) \geq p_{G^*}(x)} p_{G^*}(x) \mathrm{d}x \tag{7}$$

$$= m \int_x \left(\mathbb{1}_{p_{data}(x) < p_{G^*}(x)} p_{data}(x) + \left(1 - \mathbb{1}_{p_{data}(x) < p_{G^*}(x)} \right) p_{G^*}(x) \right) \mathrm{d}x \tag{8}$$

$$= m \int_x p_{G^*}(x) \mathrm{d}x + m \int_x \mathbb{1}_{p_{data}(x) < p_{G^*}(x)} (p_{data}(x) - p_{G^*}(x)) \mathrm{d}x \tag{9}$$

$$= m + m \int_x \mathbb{1}_{p_{data}(x) < p_{G^*}(x)} (p_{data}(x) - p_{G^*}(x)) \mathrm{d}x. \tag{10}$$

等式 10 中的第二项是非正的，因此 $V(G^*, D^*) \leq m$。

通过将生成 p_{data} 的理想生成器放入等式 4 的右侧，我们得到

$$\int_x p_{G^*}(x) D^*(x) \mathrm{d}x \leq \int_x p_{data}(x) D^*(x) \mathrm{d}x. \tag{11}$$

于是由(6)有，

$$\int_x p_{G^*}(x) D^*(x) \mathrm{d}x + \int_x p_{G^*}(x) [m - D^*(x)]^+ \mathrm{d}x \leq V(G^*, D^*) \tag{12}$$

又因为 $D^*(x) \leq m$，我们有 $m \leq V(G^*, D^*)$。

因此，$m \leq V(G^*, D^*) \leq m$，即 $V(G^*, D^*) = m$。 使用等式 10，我们看到这种情况只有在 $\int_x \mathbb{1}_{p_{data}(x) < p_G(x)} \mathrm{d}x = 0$ 时发生，这一说法为真当且仅当 $p_G = p_{data}$ 几乎无处不在（这是因为 p_{data} 和 p_G 是概率密度，详见附录 A 中的引理 2）。

定理 2。存在该系统的纳什均衡，其特征在于（a）$p_{G^*} = p_{data}$（几乎无处不在）和（b）存在常数 $\gamma \in [0, m]$ 使得 $D^*(x) = \gamma$（几乎每个地方）（见页底 1）。

[1] 这里假设没有 p_{data}(x)= 0 的区域。如果存在这样的区域，则 $D^*(x)$ 可以在[0，m]在这个地区的 x。

证明。见附件 A。 □

2.3 使用自动编码器

在我们的实验中，判别器 D 被构造为自动编码器：

$$D(x) = ||Dec(Enc(x)) - x||. \tag{13}$$

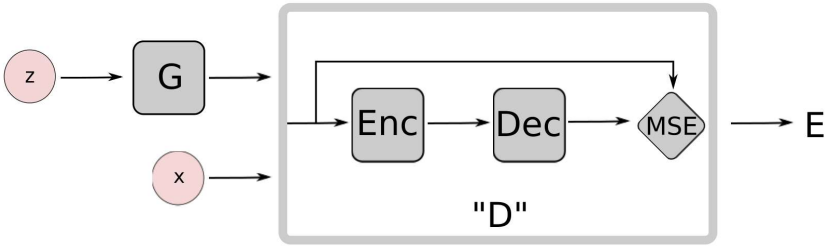

图 1：带有自动编码器的判别器的 EBGAN 架构。

具有自动编码器的鉴别器的 EBGAN 模型如图 1 所示。对于 D 的自动编码器的选择乍一看似乎是任意的，但我们假设它在概念上比二元逻辑斯谛网络更具吸引力：

- 基于重建的输出不是使用单个比特的目标信息来训练模型，而是为鉴别器提供了多样化的目标。由于二元逻辑损失，只有两个类别是可能的，因此在一个小批量中，对应于不同样本的梯度很可能远离正交。这导致低效的培练，并且减少小批量大小通常不是当前硬件的选项。另一方面，重建损失可能在小批量内产生非常不同的梯度方向，从而允许更大的小批量尺寸而不损失效率。
- 传统上，自动编码器被用于表示基于能量的模型，并且自然而然地出现。当使用一些正则化术语（参见第 2.3.1 节）进行训练时，自动编码器能够在没有监督或负面示例的情况下学习能量流形。这意味着即使在训练 EBGAN 自动编码模型来重建真实样本时，鉴别器也有助于自己发现数据流形。相反，在没有来自生成器的负样例的情况下，用二元逻辑损失训练的鉴别器变得毫无意义。

2.3.1 与常规自动编码器的连接

训练自动编码器的一个常见问题是模型可能只学习一个辨识函数，这意味着它将零能量归因于整个空间。为了避免这个问题，必须推动模型以向数据流形外部的点提供更高的能量。理论和实验结果通过规范潜在表征来解决该问题（Vincent 等人，2010; Rifai 等人，2011; MarcAurelio Ranzato 和 Chopra，2007; Kavukcuoglu 等人，2010）。这种正则化器旨在限制自动编码器的重建能力，使得它只能将低能量归因于输入点的较小部分。

我们认为，EBGAN 框架中的能量函数（鉴别器）也被视为通过生成产生对比样本的生成器来正则化，其中鉴别器应该给予高重建能量。我们进一步认为，从这个角度来看，EBGAN 框架允许更多的灵活性，因为：（i） - 正则化器（生成器）完全可以训练，而不是人工制作；（ii） - 对抗性训练范式使得产生对比样本的二元性与学习能量函数之间能够直接相互作用。

2.4 排斥正则化器

我们提出了一种“排斥正则化器”，它非常适合 EBGAN 自动编码器模型，故意保持模型不会产生聚集在p_{data}的一种或几种模式中的样本。 另一种技术“小批量判别”由 Salimans 等人开发（2016）来自同样的哲学。

实施排斥正则化程序涉及以表示层面运行的“拉出限度”（Pulling-away Term，简称 PT）。形式上，令$S \in \mathbb{R}^{s \times N}$表示从编码器输出层取得的一批样本表示。 让我们将 PT 定义为：

$$f_{PT}(S) = \frac{1}{N(N-1)} \sum_i \sum_{j \neq i} \left(\frac{S_i^\top S_j}{\|S_i\| \|S_j\|} \right)^2. \tag{14}$$

PT 在小批量上运行并尝试使成对样本表示正交化。 它的灵感来自先前的工作，显示了自动编码器相似模型中编码器的代表性能力，如 Rasmus 等（2015）和赵等人（2015 年）。 选择余弦相似度而不是欧几里德距离的基本原理是使限度限制在下方并且不变。 我们使用符号“EBGAN-PT”来指代用该限度训练的 EBGAN 自动编码器模型。 注意 PT 用于生成器损失，但不用于鉴别器损失。

3 相关工作

我们的工作主要是将 GAN 转变为基于能量的模型范围。 在这个方向上，研究对比样本的方法与 EBGAN 相关，例如使用噪声样本（Vincent 等，2010）和噪声梯度下降方法，如对比散度（Carreira-Perpinan 和 Hinton，2005）。 从 GAN 的角度出发，提出了几篇提高 GAN 训练稳定性的论文（Salimans 等，2016; Denton 等，2015; Radford 等，2015; Im 等，2016; Mathieu 等，2015）。

Kim＆Bengio（2016）提出了一种概率 GAN，并通过使用 Gibbs 分布将其投入基于能量的密度估计。与 EBGAN 不同，这个提出的框架并没有摆脱分区函数的计算挑战，因此能量函数的选择需要是可集成的。

4 实验

4.1 对 MNIST 的遗漏网格搜索

在本节中，我们研究了 EBGAN 在 GAN 上的训练稳定性，这是一个简单的 MNIST 数字生成与完全连接网络的任务。我们对基于一组架构选择和各种超参数框架上的网格进行详尽的搜索。

在形式上，我们在表 1 中指定搜索网格。我们对 EBGAN 模型施加以下限制：（i） - 对 G 和 D 使用学习率 0.001 和 Adam（Kingma＆Ba，2014）算法；（ii）-nLayerD 表示组合 Enc 和 Dec 的层总数。为简单起见，我们将 Dec 固定为一层并仅调整 Enc #layers；（iii） - 边际设定为 10 且未调整。为了分析结果，我们使用初始评分（Salimans 等，2016）作为反映生成质量的数值方法。对方法进行了一些细微的修改，使得图 2 在保持得分的原始含义的同时更加平易近人，$I' = E_x KL(p(y)||p(y|\mathbf{x}))$（见页底 2 和附录 C 中的更多细节）。简而言之，较高的I'得分意味着更好的生成质量。

直方图我们绘制了图 2 中I'得分的直方图。我们进一步从 GAN 的网格（optimD，optimG 和 lr）中分离出优化相关设置，并分别绘制每个子网格的直方图，以及 EBGAN I'得分作为参考在图 3 中，GAN 和 EBGAN 的实验数量在每个子图中都是 512。直方图显然表明 EBGAN 可以更可靠地训练。

[2] 这种形式的“初始得分”仅用于更好地分析本工作范围内的网格搜索，但不能与任何其他已发表的工作进行比较。

表 1: 网格搜索规范

Settings	Description	EBGANs	GANs
nLayerG	number of layers in G	[2, 3, 4, 5]	[2, 3, 4, 5]
nLayerD	number of layers in D	[2, 3, 4, 5]	[2, 3, 4, 5]
sizeG	number of neurons in G	[400, 800, 1600, 3200]	[400, 800, 1600, 3200]
sizeD	number of neurons in D	[128, 256, 512, 1024]	[128, 256, 512, 1024]
dropoutD	if to use dropout in D	[true, false]	[true, false]
optimD	to use Adam or SGD for D	adam	[adam, sgd]
optimG	to use Adam or SGD for G	adam	[adam, sgd]
lr	learning rate	0.001	[0.01, 0.001, 0.0001]
#experiments:	-	**512**	**6144**

从呈现最佳初始分数的配置生成的数字如图 4 所示。

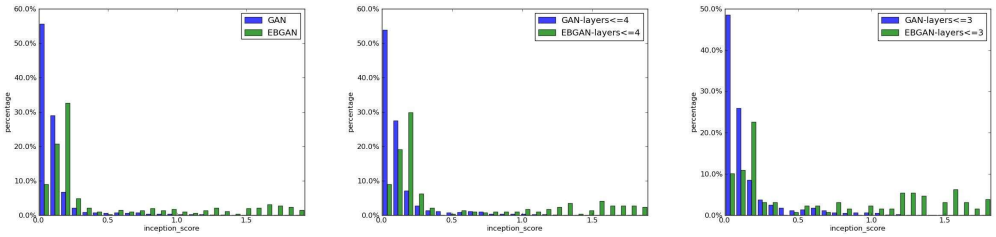

图 2 :**(建议放大 pdf 文件)** 网格搜索的初始分数的直方图。 x 轴带有初始分数 I，且 y 轴表示模型中的部分（百分比）落入某些柱（bin）中。 左（a）：EBGAN 与 GAN 的一般比较; 中（b）：EBGAN 和 GAN 均受 nLayer [GD] <= 4 的约束; 右（c）：EBGAN 和 GAN 都受 nLayer [GD] <= 3 的约束。

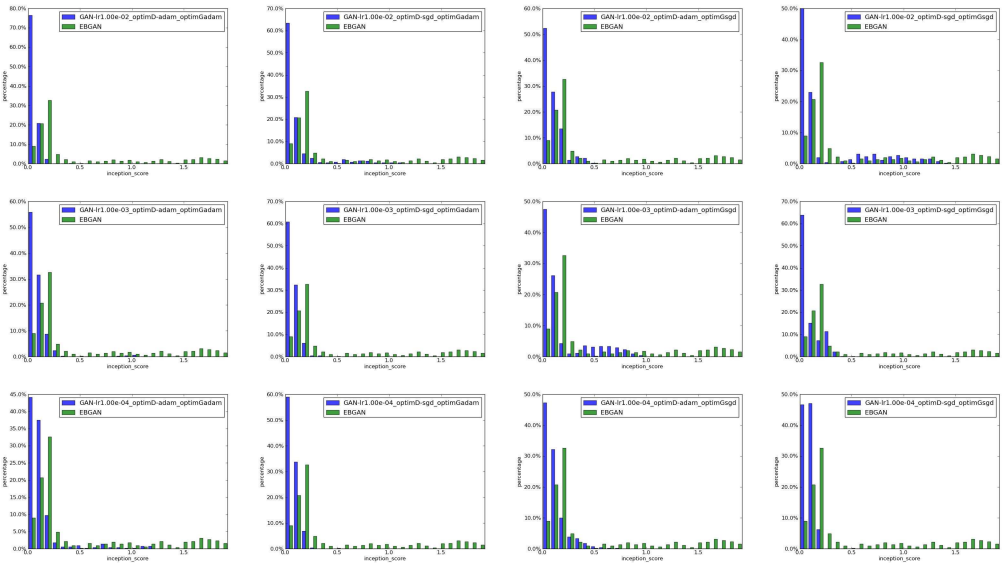

图 3 :**(建议放大 pdf 文件)** 初始分数的直方图按不同的优化组合分组，从 optimAL，optimUS 和 lr 中抽取（参见文本）。

4.2 对 MNIST 的半监督学习

我们探讨了使用 EBGAN 框架对置换不变 MNIST 进行半监督学习的潜力，共同使用 100,200 和 1000 个标签。 我们用了一个

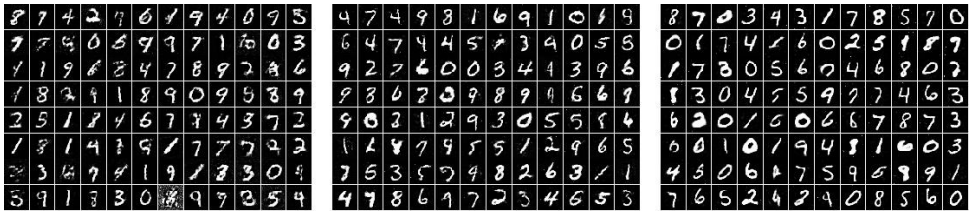

图 4：从 MNIST 上的网格搜索生成。 左（a）：最佳 GAN 模型; 中（b）：最佳 EBGAN 模型。 右（c）：最佳 EBGAN-PT 模型。

具有 EGBAN 框架（EBGAN-LN）的底层成本梯形网络（LN）（Rasmus 等，2015）。梯形网络可以归类为基于能量的模型，该模型由前馈和反馈层次结构构成，这些层次结构由耦合两个路径的阶段式横向连接提供动力。

我们发现在启用 EBGAN 框架进行半监督学习时至关重要的一种技术是逐渐衰减等式 1 的边际值 m。其背后的基本原理是当 p_G 接近数据流形时让鉴别器惩罚生成器。人们可以想到对比样本被精确地固定在数据流形上的极端情况，使得它们“不再具有相关性”。当 m = 0 并且 EBGAN-LN 模型回退到正常的梯形网络时，会发生这种最终状态。在 GAN 或 EBGAN 框架中使用鉴别器的非衰减动力学的不可取性也由定理 2 表示：在收敛时，鉴别器反映了平坦的能量表面。然而，我们认为学习 EBGAN-LN 模型的轨迹确实通过让它看到对比样本来为 LN（鉴别器）提供更多信息。然而，避免上述不受欢迎程度的最佳方法是确保在达到纳什均衡时 m 已经衰减到 0。在我们的实验中通过超参数搜索找到边际衰减时间表（附录 D 中的技术细节）。

从表 2 可以看出，将底层成本 LN 定位到 EBGAN 框架中可以显著提高 LN 本身的性能。我们假设在 EBGAN 框架的范围内，迭代地将由生成器产生的对抗性对比样本馈送到能量函数中作为有效的正则化器;对比样本可以被认为是数据集的扩展，为分类器提供更多信息。我们注意到 Rasmus 等报道的结果之间存在差异（2015 年）和 Pezeshki 等人（2015），所以我们报告两个结果以及我们自己的梯形网络实施相同的设置。具体的实验设置和分析见附录 D.

表 2：LN 底层成本模型与其在 EB-MNIST 半监督任务中的 EBGAN 扩展的比较。 注意结果是错误率（以%表示）并在 15 个不同的随机种子上取平均值。

model	100	200	1000
LN bottom-layer-cost, reported in Pezeshki et al. (2015)	1.69±0.18	-	1.05±0.02
LN bottom-layer-cost, reported in Rasmus et al. (2015)	1.09±0.32	-	0.90±0.05
LN bottom-layer-cost, reproduced in this work (see appendix D)	1.36±0.21	1.24±0.09	1.04±0.06
LN bottom-layer-cost within EBGAN framework	**1.04±0.12**	**0.99±0.12**	**0.89±0.04**
Relative percentage improvement	23.5%	20.2%	14.4%

4.3 LSUN 与 CELEBA

我们应用具有深度卷积结构的 EBGAN 框架来生成 64 × 64 个 RGB 图像，这是一个更现实的任务，使用 LSUN 卧室数据集（Yu et al. 2015）和大规模人脸数据集 CelebA（Liu 等人，2015 年）。为了将 EBGAN 与 DCGAN 进行比较（Rad-ford 等，2015），我们在相同的配置下训练 DCGAN 模型，并在图 5 和图 6 中与 EBGAN 模型并排显示。具体设置列出在附录 C 中。

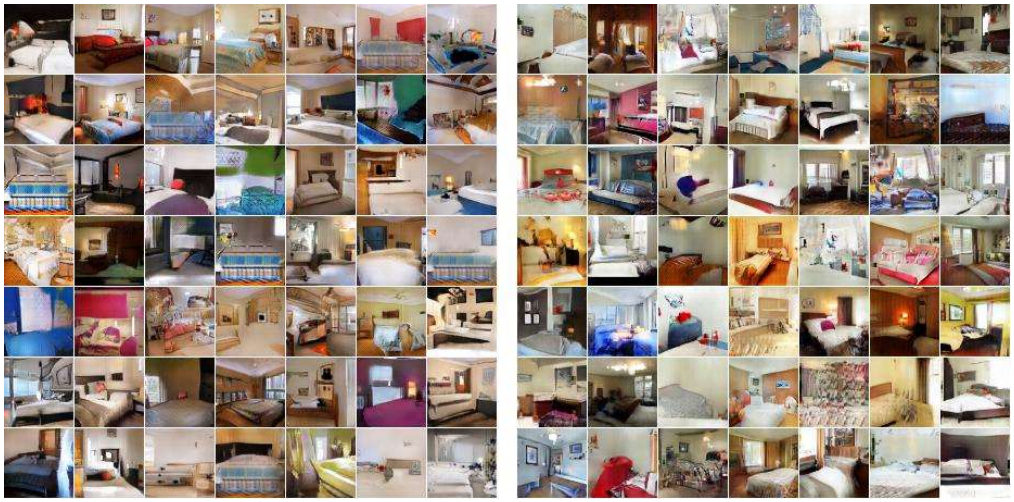

图 5：LSUN 卧室数据集的生成。 左（a）： DCGAN 生成。 右（b）： EBGAN-PT 生成。

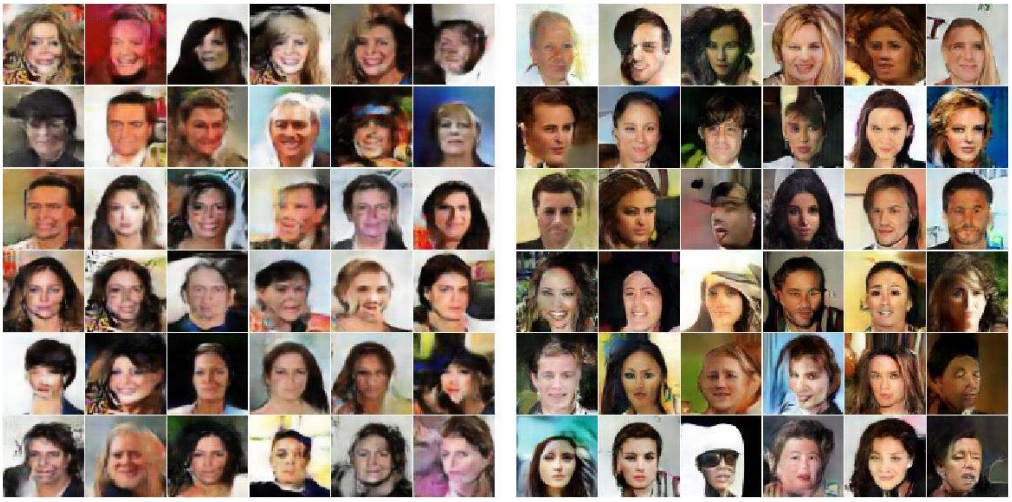

图 6：CelebA 数据集的生成。 左（a）： DCGAN 生成。 右（b）： EBGAN-PT 生成。

4.4 IMAGENET

最后，我们训练 EBGAN 在 ImageNet 上生成高分辨率图像（Russakovsky 等，2015）。与我们迄今为止进行过实验的数据集相比，ImageNet 提供了更大和更广阔的空间，因此通过生成模型对数据分布进行建模变得非常具有挑战性。我们设计了一个实验来生成 128 × 128 的图像，这些图像在完整的 ImageNet-1k 数据集上进行了训练，其中包含来自 1000 个不同类别的大约 130 万个图像。我们还训练了一个网络，使用 Vinyals 等人提供的 wordNet IDs，在 ImageNet 的狗品种子集上生成大小为 256 × 256 的图像（2016）。结果显示在图 7 和图 8 中。尽管难以在高分辨率级别生成图像，但我们观察到 EBGAN 能够了解对象出现在前景中的事实，以及类似草的各种背景组件纹理，地平线下的海，水中的镜像山，建筑物等。此外，我们的 256 × 256 狗品种，虽然远非现实，确实反映了一些关于狗的外观，如他们的身体，毛皮和眼睛的知识。

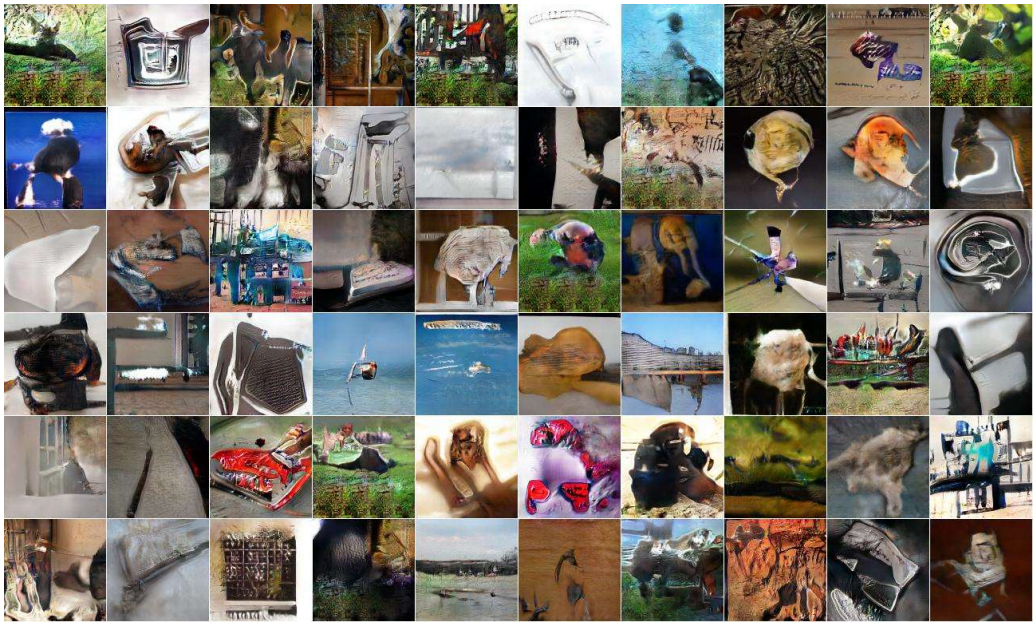

图 7：使用 EBGAN-PT 的 ImageNet 128 × 128 生成图。

图 8：使用 EBGAN-PT 的 ImageNet 256 × 256 生成图。

5 展望

我们桥接了两类无监督学习方法 - GAN 和自动编码器 - 并从替代的能量角度重新审视 GAN 框架。 EBGAN 显示出更好的收敛模式和可扩展性，可生成高分辨率图像。 LeCun 等人提出了一系列基于能量的损失函数（2006）可以很容易地纳入 EBGAN 框架。 对于未来的工作，条件约束集（Denton 等，2015; Mathieu 等，2015）是一个设置的设置。 我们希望未来的研究能够从基于能量的角度更加关注更广泛的 GAN 观点。

致谢

我们感谢 Emily Denton，Soumith Chitala，Arthur Szlam，Marc'Aurelio Ranzato，Pablo Sprech-mann，Ross Goroshin 和 Ruoyu Sun 进行了富有成效的讨论。 我们也感谢 Emily Denton 和 Tian Jiang 对手稿的帮助。

参考文献

Carreira-Perpinan, Miguel A and Hinton, Geoffrey. On contrastive divergence learning. In AISTATS, volume 10, pp. 33–40. Citeseer, 2005.

Denton, Emily L, Chintala, Soumith, Fergus, Rob, et al. Deep generative image models using a laplacian pyramid of adversarial networks. In Advances in neural information processing systems, pp. 1486–1494, 2015.

Goodfellow, Ian, Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron, and Bengio, Yoshua. Generative adversarial nets. In Advances in Neural Information Processing Systems, pp. 2672–2680, 2014.

Im, Daniel Jiwoong, Kim, Chris Dongjoo, Jiang, Hui, and Memisevic, Roland. Generating images with recur-rent adversarial networks. arXiv preprint arXiv:1602.05110, 2016.

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

Kavukcuoglu, Koray, Sermanet, Pierre, Boureau, Y-Lan, Gregor, Karol, Mathieu, Michael,¨ and Cun, Yann L. Learning convolutional feature hierarchies for visual recognition. In Advances in neural information pro-cessing systems, pp. 1090–1098, 2010.

Kim, Taesup and Bengio, Yoshua. Deep directed generative models with energy-based probability estimation. arXiv preprint arXiv:1606.03439, 2016.

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

LeCun, Yann, Chopra, Sumit, and Hadsell, Raia. A tutorial on energy-based learning. 2006.

Liu, Ziwei, Luo, Ping, Wang, Xiaogang, and Tang, Xiaoou. Deep learning face attributes in the wild. In Proceedings of the IEEE International Conference on Computer Vision, pp. 3730–3738, 2015.

MarcAurelio Ranzato, Christopher Poultney and Chopra, Sumit. Efficient learning of sparse representations with an energy-based model. 2007.

Mathieu, Michael, Couprie, Camille, and LeCun, Yann. Deep multi-scale video prediction beyond mean square error. arXiv preprint arXiv:1511.05440, 2015.

Pezeshki, Mohammad, Fan, Linxi, Brakel, Philemon, Courville, Aaron, and Bengio, Yoshua. Deconstructing the ladder network architecture. arXiv preprint arXiv:1511.06430, 2015.

Radford, Alec, Metz, Luke, and Chintala, Soumith. Unsupervised representation learning with deep convolu-tional generative adversarial networks. arXiv preprint arXiv:1511.06434, 2015.

Ranzato, Marc'Aurelio, Boureau, Y-Lan, Chopra, Sumit, and LeCun, Yann. A unified energy-based framework for unsupervised learning. In Proc. Conference on AI and Statistics (AI-Stats), 2007.

Rasmus, Antti, Berglund, Mathias, Honkala, Mikko, Valpola, Harri, and Raiko, Tapani. Semi-supervised learning with ladder networks. In Advances in Neural Information Processing Systems, pp. 3546–3554, 2015.

Rifai, Salah, Vincent, Pascal, Muller, Xavier, Glorot, Xavier, and Bengio, Yoshua. Contractive auto-encoders: Explicit invariance during feature extraction. In Proceedings of the 28th international conference on machine learning (ICML-11), pp. 833–840, 2011.

Russakovsky, Olga, Deng, Jia, Su, Hao, Krause, Jonathan, Satheesh, Sanjeev, Ma, Sean, Huang, Zhiheng, Karpathy, Andrej, Khosla, Aditya, Bernstein, Michael, Berg, Alexander C., and Fei-Fei, Li. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV), 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Salimans, Tim, Goodfellow, Ian, Zaremba, Wojciech, Cheung, Vicki, Radford, Alec, and Chen, Xi. Improved techniques for training gans. arXiv preprint arXiv:1606.03498, 2016.

Vincent, Pascal, Larochelle, Hugo, Lajoie, Isabelle, Bengio, Yoshua, and Manzagol, Pierre-Antoine. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. Journal of Machine Learning Research, 11(Dec):3371–3408, 2010.

Vinyals, Oriol, Blundell, Charles, Lillicrap, Timothy, Kavukcuoglu, Koray, and Wierstra, Daan. Matching networks for one shot learning. arXiv preprint arXiv:1606.04080, 2016.

Yu, Fisher, Seff, Ari, Zhang, Yinda, Song, Shuran, Funkhouser, Thomas, and Xiao, Jianxiong. Lsun: Con-struction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint arXiv:1506.03365, 2015.

Zhao, Junbo, Mathieu, Michael, Goroshin, Ross, and Lecun, Yann. Stacked what-where auto-encoders. arXiv preprint arXiv:1506.02351, 2015.

A 附录：第 2.2 节的技术要点

引理 1。设$a, b \geq 0, \varphi(y) = ay + b[m - y]^+$。 φ 在$[0, +\infty)$上的最小值存在，并且如果 a < b 则最小值在 m 处取得，否则在 0 处取得（否则最小值可能不唯一）。

证明。函数φ定义在$[0, +\infty)$，其导数定义在$[0, +\infty)\backslash\{m\}$ 并且如果$y \in [0, m)$ 则$\varphi'(y) = a - b$，如果$y \in (m, +\infty)$ 则$\varphi'(y) = a$。

所以当 a < b 时，函数在[0，m）上减少并且在（m，+∞）上增加。 由于它是连续的，因此它的最小值为 m。 如果 a = 0 或 a - b = 0，它可能不是唯一的。

另一方面，如果 a ≥ b 函数φ在[0，+∞）上增加，所以 0 是最小值。

引理 2。 如果 p 和 q 是概率密度, 那么$\int_x \mathbb{1}_{p(x)<q(x)}\mathrm{d}x = 0$ 当且仅当 $\int_x \mathbb{1}_{p(x)\neq q(x)}\mathrm{d}x = 0$

证明。假设$\int_x \mathbb{1}_{p(x)<q(x)}\mathrm{d}x = 0$，则

$$\begin{aligned}
& \int_x \mathbb{1}_{p(x)>q(x)}(p(x) - q(x))\mathrm{d}x && (15)\\
= & \int_x (1 - \mathbb{1}_{p(x)\leq q(x)})(p(x) - q(x))\mathrm{d}x && (16)\\
= & \int_x p(x)\mathrm{d}x - \int_x q(x)\mathrm{d}x + \int_x \mathbb{1}_{p(x)\leq q(x)}(p(x) - q(x))\mathrm{d}x && (17)\\
= & 1 - 1 + \int_x \left(\mathbb{1}_{p(x)<q(x)} + \mathbb{1}_{p(x)=q(x)}\right)(p(x) - q(x))\mathrm{d}x && (18)\\
= & \int_x \mathbb{1}_{p(x)<q(x)}(p(x) - q(x))\mathrm{d}x + \int_x \mathbb{1}_{p(x)=q(x)}(p(x) - q(x))\mathrm{d}x && (19)\\
= & 0 + 0 = 0 && (20)
\end{aligned}$$

所以$\int_x \mathbb{1}_{p(x)>q(x)}(p(x) - q(x))\mathrm{d}x = 0$并且因为积分中的项总是非负的，所以对几乎所有的 x 都有$\mathbb{1}_{p(x)>q(x)}(p(x) - q(x)) = 0$。并且$p(x) - q(x) = 0$意味着$\mathbb{1}_{p(x)>q(x)} = 0$，所以$\mathbb{1}_{p(x)>q(x)} = 0$几乎无处不在。因此$\int_x \mathbb{1}_{p(x)>q(x)}\mathrm{d}x = 0$在给定假设下完成了证明。

（第三页的）**定理 2 证明** 充分条件是显而易见的。 G^* 的必要条件来自定理 1，$D^*(x) \leq m$的必要条件来自定理 1 的证明。

现在让我们假设G^* 几乎无处不连续，并且发现矛盾。 如果不是，则存在常数 C 和非零测量的集合 S，使得$\forall x \in \mathcal{S}, D^*(x) \leq C$ 和$\forall x \notin \mathcal{S}, D^*(X) > C$。另外我们可以选择 S 使得存在非零测量的子集$\mathcal{S}' \subset \mathcal{S}$，使得在$\mathcal{S}'$ 上$p_{data}(x) > 0$（因为在脚注中的假设）。 我们可以构建一个生成器G_0，使得在 S 上$p_{G_0}(x) \leq p_{data}(x)$，而在$\mathcal{S}'$ 上$p_{G_0}(x) < p_{data}(x)$。 我们计算

$$\begin{aligned}
U(G^*, D^*) - U(G_0, D^*) &= \int_x (p_{data} - p_{G_0})D^*(x)\mathrm{d}x && (21)\\
&= \int_x (p_{data} - p_{G_0})(D^*(x) - C)\mathrm{d}x && (22)\\
&= \int_{\mathcal{S}} (p_{data} - p_{G_0})(D^*(x) - C)\mathrm{d}x + \int_{\mathcal{R}^N\backslash\mathcal{S}} (p_{data} - p_{G_0})(D^*(x) - C)\mathrm{d}x && (23)\\
&> 0 && (24)
\end{aligned}$$

这违反了等式 4。

B 附录：关于 GANs 和基于能量的学习的更多解释

GANs 的两种解释

GAN 可以用两种互补的方式解释。在第一种解释中，关键部分是生成器，鉴别器起到可训练目标函数的作用。让我们想象一下，数据存在于多方面。直到生成器产生被识别为在流形上的样本，它才会得到一个梯度，指示如何修改其输出以使其接近流形。在这种情况下，鉴别器用于在发生器产生在流形外部的样本时惩罚生成器。这可以被理解为用传统的监督学习来训练具有一组可能的期望输出（例如，流形）的生成器而不是单个期望输出的方式。

对于第二种解释，关键部分是鉴别器，并且仅生成生成器以产生对比样本。我们通过迭代和交互式馈送对比样本表明，生成器在 4.2 节中增强了鉴别器（例如梯形网络）的半监督学习性能。

C 附录：实验设置

有关网格搜索的更多详细信息

为了训练网格搜索的 EBGAN 和 GAN，我们使用以下设置：

- 除了生成器输出层和鉴别器输入层之外，在每个权重层之后应用批量归一化（Ioffe 和 Szegedy，2015）（Radford 等人，2015）。
- 训练图像缩放到范围[-1,1]。 相应地，生成器输出层之后是 Tanh 函数。
- ReLU 用作非线性函数。
- 初始化：D 中的权重来自 N（0，0.002），G 中来自 N（0，0.02）。偏差初始化为 0。

我们通过计算初始得分的修改版本来评估网格搜索中的模型，$I' = E_x KL(p(y)||p(y|\mathbf{x}))$，其中 x 表示生成的样本，y 是由 MNIST 分类器预测的标签。 使用整个 MNIST 训练集离线训练。 对其原始形式进行了两项主要修改：（i）- 我们交换了分配对的顺序；（ii）- 我们省略了 $e^{(\cdot)}$ 操作。 修改后的分数浓缩了图 2 和图 3 中的直方图。值得注意的是，尽管我们继承了 Salimans 等人的名称“起始分数”（2016），评估与在 ImageNet 数据集上训练的“初始”模型无关。 分类器是在 MNIST 上训练的常规 3 层 ConvNet。

图 4 中显示的生成图是网格搜索中最好的 GAN 或 EBGAN（获得最佳 I' 分数）。 他们的配置是：

- 图 4(a): nLayerG=5, nLayerD=2, sizeG=1600, sizeD=1024, dropoutD=0, optimD=SGD, optimG=SGD, lr=0.01.
- 图 4(b): nLayerG=5, nLayerD=2, sizeG=800, sizeD=1024, dropoutD=0, optimD=ADAM, optimG=ADAM, lr=0.001, margin=10.
- 图 4(c): same as (b), with P T = 0.1。

LSUN 与 CELEBA

我们使用类似于 DCGAN 的深度卷积生成器和用于鉴别器的深度卷积自动编码器。 自动编码器由前馈路径中的跨步卷积模块和反馈路径中的分数跨度卷积模块组成。 我们将使用上采样或 switches-unpooling（Zhao et al. 2015）用于未来的研究。 我们也遵循 Radford 等人提出的指导（2015 年）训练 EBGAN。 深度自动编码器的配置是：

- Encoder: (64)4c2s-(128)4c2s-(256)4c2s
- Decoder: (128)4c2s-(64)4c2s-(3)4c2s

其中“(64)4c2s”表示卷积/反卷积层，具有 64 个输出特征映射，内核大小 4 具有步幅 2。对于 LSUN，边际 m 设置为 80，对于 CelebA，设置为 20。

IMAGENET

我们在 128 × 128 和 256 × 256 个实验中建立了更深的模型，与 4.3 节类似，

- 128 × 128 model:
  - Generator: (1024)4c-(512)4c2s-(256)4c2s-(128)4c2s-(64)4c2s-(64)4c2s-(3)3c
  - Noise #planes: 100-64-32-16-8-4
  - Encoder: (64)4c2s-(128)4c2s-(256)4c2s-(512)4c2s
  - Decoder: (256)4c2s-(128)4c2s-(64)4c2s-(3)4c2s
  - Margin: 40
- 256 × 256 model:
  - Generator: (2048)4c-(1024)4c2s-(512)4c2s-(256)4c2s-(128)4c2s-(64)4c2s-(64)4c2s-(3)3c
  - Noise #planes: 100-64-32-16-8-4-2
  - Encoder: (64)4c2s-(128)4c2s-(256)4c2s-(512)4c2s
  - Decoder: (256)4c2s-(128)4c2s-(64)4c2s-(3)4c2s
  - Margin: 80

请注意，我们将噪声馈送到生成器的每个层中，其中每个噪声分量被初始化为 4D 张量并与特征空间中的当前特征映射连接。 Salimans 等人也采用了这种策略（2016）。

D 附录：半监督学习实验设置

基线模型

如 4.2 节所述，我们选择了底层成本梯形网作为我们的基线模型。具体而言，我们利用两篇论文中报道的相同架构（Rasmus 等，2015; Pezeshki 等，2015）;即一个大小为 784-1000-500-250-250-250 的全连接网络，在每个线性层之后具有批量归一化和 ReLU。为了获得强大的基线，我们使用集合$\{\frac{5000}{784}, \frac{2000}{784}, \frac{1000}{784}, \frac{500}{784}\}$的值调整重建损失的权重，同时将分类损失的权重固定为 1。同时，我们还调整了学习率值{ 0.002,0.001,0.0005,0.0002,0.0001 }。我们采用 Adam 作为优化器，β_1设置为 0.5。小批量大小设置为 100。所有实验完成 120,000 步。我们使用与已发表论文中相同的学习率衰减机制 - 从总步骤的三分之二（即步骤＃80,000）开始，将学习率线性衰减为 0。4.2 节中报告的结果由最佳调整设置：$\lambda_{L2} = \frac{1000}{784}, lr = 0.0002$。

EBGAN-LN 模型

我们将相同的梯形网络架构放入我们的 EBGAN 框架中，并以与训练 EBGAN 自动编码器模型相同的方式训练此 EBGAN-LN 模型。 对于技术细节，我们从边际值 16 开始训练 EBGAN-LN 模型，并在前 60,000 步中逐渐将其衰减到 0。 到那时，我们发现实际图像的重建损失已经很低并且达到了体系结构的限制（梯形网络本身）；此外，生成的图像表现出良好的质量（如图 10 所示）。 此后，我们关闭了对生成器的训练，但继续训练鉴别器另外 120,000 步。 我们将鉴别器的初始学习率设置为 0.0005，将生成器设置为 0.00025。 另一个设置与最佳基线 LN 模型保持一致。 学习率下降从步骤＃120,000（也是总步数的三分之二）开始。

其他细节

- 请注意，我们在 EBGAN-LN 实验中使用了 MNIST 数据集的 28 × 28 版本（未填充）。对于 EBGAN 自动编码器网格搜索实验，我们使用零填充版本，即大小为 32 × 32。由于零填充，未发现任何显著差异。
- 我们通常采用 EBGAN 自动编码器模型中输入和重建之间差异的损失项的 ℓ_2 范数，正如 2.1 节中正式编写的那样。 然而，对于 EBGAN-LN 实验，我们遵循梯形网络的原始实现，使用了 ℓ_2 损失的 vanilla 形式。
- 借用 Salimans 等人（2016），采用批量归一化而没有学习参数γ，但仅采用偏差项β。目前仍然不知道这种伎俩是否会以某种不可忽视的方式影响学习，因此这可能使我们的基线模型不能严格复制 Rasmus 等人发表的模型(2015 年)和 Pezeshki 等人(2015 年)。

E 附录：设定良好能源边际值的提示

从理论和实验的角度来看，在 EBGAN 的框架内设置适当的能量边际值 m 至关重要。 在此我们提供一些提示：

- 深入研究由等式 1 得出的鉴别器损失的公式，我们建议在它的两个关于真实样本和假样本的项之间的数值平衡。第二项显然受[0，m]限制（假设能量函数 D（x）是非负的）。使第一项限制在相似的范围内是值得的。理论上，第一项的上限基本上由（i）- D 的容量;（ii）- 数据集的复杂性决定。
- 实际上，对于 EBGAN 自动编码器模型，可以在实际样本数据集上单独运行 D（自动编码器）并监视损失。当它收敛时，相应的损失意味着对 D 的这种设置能够适合数据集的程度的粗略限制是多少。这通常意味着在 m 上进行超参数搜索的良好开端。
- m 过大导致训练不稳定/困难，而 m 太小则容易出现模式下降问题。 m 的这种性质如图 9 所示。
- 正如我们在附录 D 中介绍的那样，一种成功的技术是从很大的 m 开始，并在训练过程中逐渐衰减到 0。与 Salimans 等人提出的特征匹配半监督学习技术不同（2016），我们在图 10 中表明，EBGAN-LN 模型不仅实现了良好的半监督学习效果，而且还产生了令人满意的生成图。

从实际的实验技巧中抽象出来，第 2.2 节对 EBGAN 的理论认识也为设定可行的 m 提供了一些见解。 例如，如定理 2 所暗示的，设置大的 m 导致令 $D^*(x)$ 可以收敛的γ范围更宽。 不稳定可能在γ过大之后出现，因为它会产生两个指向相反方向的强梯度，从损失 1（即公式（1）处）开始，将需要更多挑剔的优化设置。

F 附录：更多生成图

LSUN 增强版本训练

对于 LSUN 卧室数据集，除了对整个图像进行实验外，我们还通过裁剪补丁来训练基于数据集增强的 EBGAN 自动编码器模型。 所有贴片的尺寸均为 64 × 64，并从 96 × 96 原始图像中裁剪。 这一生成图如图 11 所示。

EBGANS 和 EBGAN-PTS 的比较

为了进一步说明拉出项（PT）如何影响 EBGAN 自动编码器模型训练，我们选择了 LSUN 卧室数据集的全图像和增强补丁版本，以及 CelebA 数据集进行进一步的实验。 EBGAN 和 EBGAN-PT 生成图的比较如图 12，图 13 和图 14 所示。注意

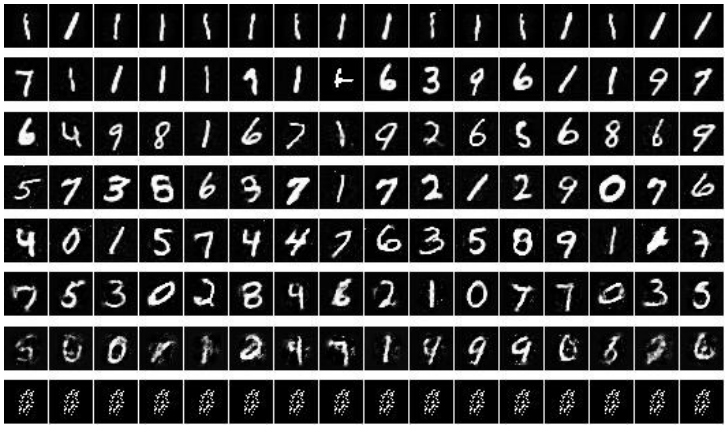

图 9：使用不同 m 设置训练的 EBGAN 自动编码器模型的生成图。 从上到下，m 分别设置为 1,2,4,6,8,12,16,32。其余设置为 nLayerG = 5，nLayerD = 2，sizeG = 1600，sizeD = 1024，dropoutD = 0，optimD = ADAM，optimG = ADAM，lr = 0.001。

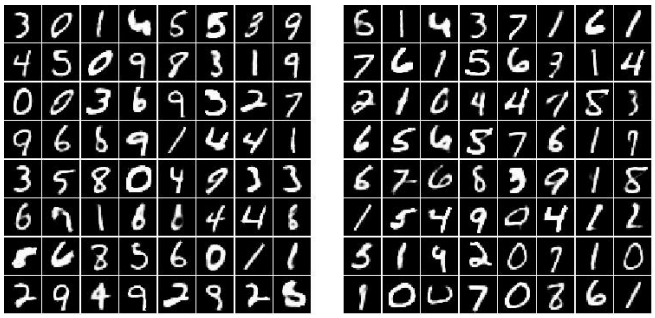

图 10：EBGAN-LN 模型的生成图。 显示的生成图通过附录 D 中描述的相同实验设置获得，具有不同的随机种子。正如我们之前提到的，我们在 EBGAN-LN 实验中使用了未填充版本的 MNIST 数据集（大小 28 × 28）。

所有比较对采用与 4.3 节相同的架构和超参数设置。 PT 上的损失权重设置为 0.1。

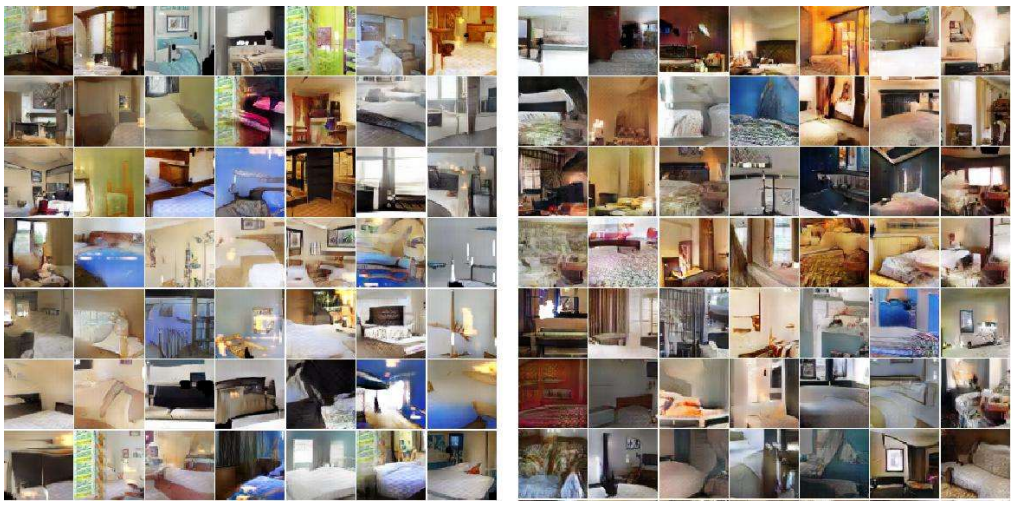

图 11：从 LSUN 卧室数据集的增强补丁版本的生成图。 左（a）：DC-GAN 生成。 右（b）：EBGAN-PT 生成。

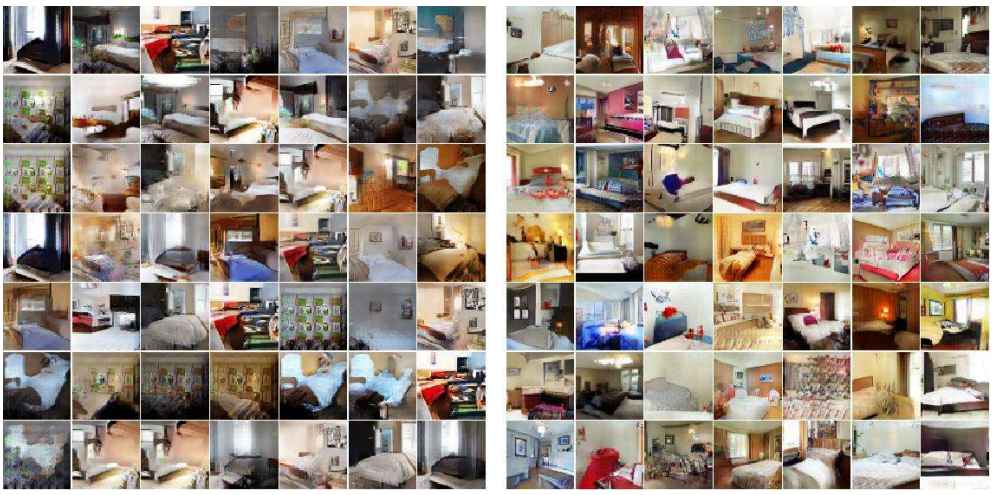

图 12：从 LSUN 卧室数据集的整个图像版本的生成图。 左（a）：EBGAN。 右（b）：EBGAN-PT。

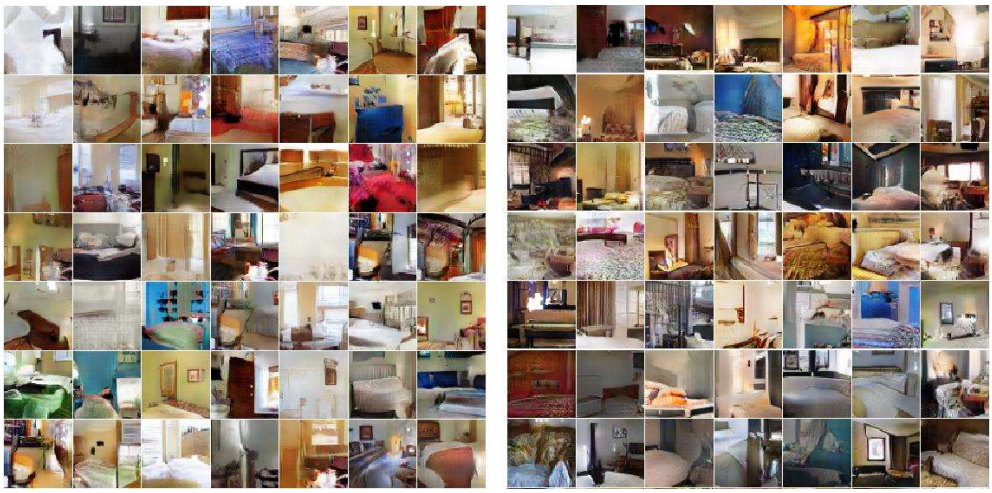

图 13：从 LSUN 卧室数据集的增强补丁版本的生成图。 左（a）： EBGAN。 右（b）： EBGAN-PT。

图 14：CelebA 数据集的生成图。 左（a）： EBGAN。 右（b）： EBGAN-PT。