

Improved Techniques for Training GANs

notes written by h1astro

1. 结构改进和训练技巧
2. 用于半监督学习
3. 生成图像的质量评价
4. 代码实现

核心要点：

提出了一系列新的GAN结构和训练方式

进行了半监督学习和图像生成相关实验

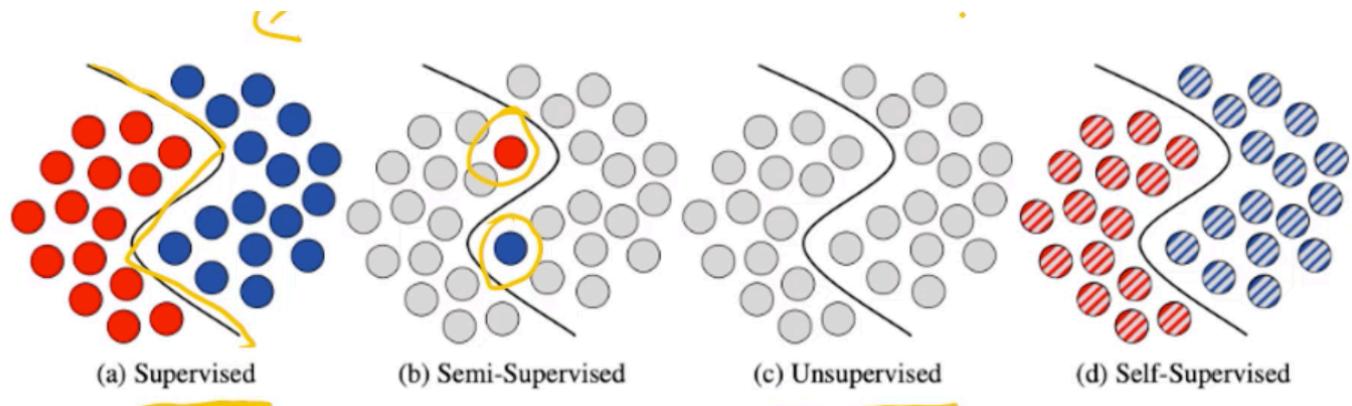
新的技术框架在MNIST、CIFAR-10和SVHN的半监督分类中取得了良好效果

通过视觉图灵测试证明，生成的图像同真实图像已难以区分

在ImageNet上训练，模型学习到了原图的显著特征

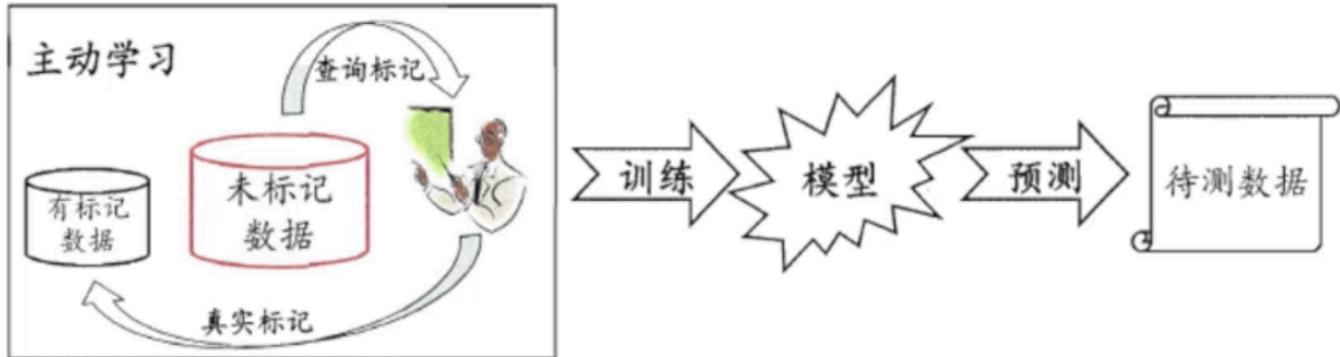
数据标签

- 监督学习：数据有完整的标签
- 无监督学习：没有数据标签
 - 自监督学习：自动生成标签
- 半监督学习：少量数据拥有完整标签
- 弱监督学习：数据有不完整的标签
- 强化学习：一开始没有标签，从环境中逐渐生成标签



半监督学习：

流形假设：将高维数据嵌入到低维流形中，当两个样例位于低维流形中的一个小局部邻域内时，具有相似的类标签



研究意义：

在DCGAN的基础上，通过多种正则化手段提升收敛性

综合质量和多样性，给出了一条评价GAN生成效果的新路径

展示了把GAN生成图像，作为其它图像任务训练集的可行性

精读

收敛性问题

GAN网络的目的是在高位非凸的参数空间中，找到一个价值函数的纳什均衡点

使用梯度下降来优化GAN网络，只能得到较低的损失，不能找到真正的纳什均衡

例如，一个网络修改 x 来最小化 xy ，另一个网络修改 y 来最小化 $-xy$ ，使用梯度下降进行优化，结果进入一个稳定的轨道中，并不会收敛到 $(0, 0)$ 点。

特征匹配

- 判别器的隐层中，经过训练逐渐得到了可用于区分真假样本的特征
- 生成器生成样本的过程中，如果也生成近似这些特征的样本，就可以更好的近似真实样本
- 尽管不能保证到达均衡点，不过收敛的稳定性应该有提高
- 用 $f(x)$ 表示判别器隐层中的特征，生成器新的目标函数被定义为：

$$\|E_{x \sim p_{data}} f(x) - E_{z \sim p_z(z)} f(G(z))\|_2^2 \text{ ---- L2距离，在判别器中间层也要求一样}$$

小批量判别

- 没有一个机制保证生成器需要生成不一样的数据
- 当模式崩溃即将发生时，判别器中许多相似点的梯度会指向一个相近的方向
- 计算判别器中某一层特征中，同一个batch各样本特征间的差异，来作为下一层的额外输入
- 能够非常快速地生成视觉上吸引人的样本，并且它的效果优于**特征匹配**
- 如果使用第5节中描述的半监督学习方法获得强分类器，则**特征匹配**比本方法效果更好

历史平均

- 在生成网络和判别网络的损失函数中添加一个历史参数的平均项，这个项在网络训练过程中也会得到更新
- $\theta[i]$ 是过去第*i*时刻，模型的所有参数值
- 加入历史平均项后，梯度就不容易进入稳定的轨道，能够继续朝着均衡点前进

$$\|\theta - \frac{1}{t} \sum_{i=1}^t \theta[i]\|^2$$

单向的标签平滑

- 判别器的目标函数中正负样本的目标输出 α 和 β 不再是1和0，而是接近它们的数字，比如0.9和0.1，则优化目标变为 $D(X)$
- 最近的研究显示，这项技术可以增强分类网络对于对抗样本的鲁棒性
- 在 p_{data} 接近0，而 p_{model} 很大时， $D(X)$ 的梯度也接近0，无法优化
- 只将正样本的标签做平滑，负样本的标签仍然设为0

$$D(x) = \frac{\alpha p_{data}(x) + \beta p_{model}(x)}{p_{data}(x) + p_{model}(x)}$$

虚拟批归一化

- batch normalization在DCGAN中被应用，能大大减少GAN优化的难度
- 但BN层有个缺点，会使生成的同一个batch中，每张图片之间存在关联（如一个batch中的图片有很多绿色）
- 为解决这个问题，训练开始前先选择一个固定的reference batch，每次算出这个特定的batch的均值和方差，再用它们对训练中的batch进行normalization
- 缺点时需要进行两次前向传播，增加了计算成本，所以只在**生成器**上应用

图像质量评价

人工评价

使用Amazon Mechanical Turk, 即亚马逊众包平台进行人工标注

将真实图片和生成图片参杂在一起, 标注着需要逐个指出给定图像是真实的还是生成的

当给标注者提供标注反馈时, 结果会发生巨大变化; 通过学习这些反馈, 标注者能够更好地指出生成图像中的缺陷, 从而更倾向于把图像标记为生成的

Inception Score

提出了一种自动评估样本的方法, 评估结果与人类的评估高度相关

使用Inception模型, 生成图片 x 为输入, 以 x 的推断类标签概率 $p(y|x)$ 为输出

单个样本的输出分布应该为低熵, 即高预测置信度, 好样本应该包含明确有意义的目标物体

所有样本的输出整体分布应该都为高熵, 也就是说, 所有的 x 应该尽量分布于不同的类, 而不是属于同一类。

因此Inception score定义为: $\exp(E_x KL(p(y|x)||p(y)))$

$$\exp\left(\frac{1}{N} \sum_{i=1}^N D_{KL}(p(y|x^{(i)})||\hat{p}(y))\right)$$

半监督学习

- 普通的分类任务属于有监督学习, 模型通过最小化交叉上损失, 获得最优的网络参数
- 在分类网络的基础上, 加入GAN的生成数据, 就可以把有监督学习转变成半监督学习
- 整个系统将误差函数拆开, 共有三种误差:

对于训练集中的有标签样本, 考察估计的标签是否正确。即, 计算分类为相应的概率

$$L_{supervised} = -\mathbb{E}_{x,y \sim p_{data}} \log p_{model}(y|x, y < K+1)$$

对于训练集中的无标签样本, 考察是否估计为“真”。即, 计算不估计为K+1类的概率

$$L_{unlabel} = -\mathbb{E}_{x,y \sim p_{data}} [\log(1 - p_{model}(y = K+1|x))]$$

对于生成器产生的伪样本, 考察是否估计为“伪”。即, 计算估计为K+1类的概率

$$L_{fake} = -\mathbb{E}_{x \sim p_G(x)} \log [p_{model}(y = K+1|x)]$$

判别器: $L_D = L_{label} + \frac{w}{2}(L_{unlabel} + L_{fake})$

生成器: $L_G = -L_{fake}$

- 除了在半监督学习中实现最先进的结果之外, 经过主观评测发现, 使用半监督学习训练出的生成器, 该生成的图像相比于原始生成器也得到了很大的提升
- 说明本方法的视觉系统, 对图像类别的理解和对图像质量的理解存在显著相关性, 也佐证了使用Inception score 来评价生成图像质量的合理性
- 这个关联性可以用来理解迁移学习的作用, 并且存在广泛的应用前景

实验结果

MNIST

- 分别使用20、50、100和200个带标签图像进行训练，剩余的训练图像则没有标签
- 结果子啊10个带标签数据的随机子集上取平均值，每个子集中，保证各个类别的数量均衡
- 网络各有5个隐层，使用了权重归一化，并且将高斯噪声加入到判别器每一层的输出
- 使用feature matching 的模型，生成的图像质量一般
- 使用minibatch 判别的模型，生成的图像质量非常好，经众包标注和内部人员评测，已基本无法与真实数据相区分

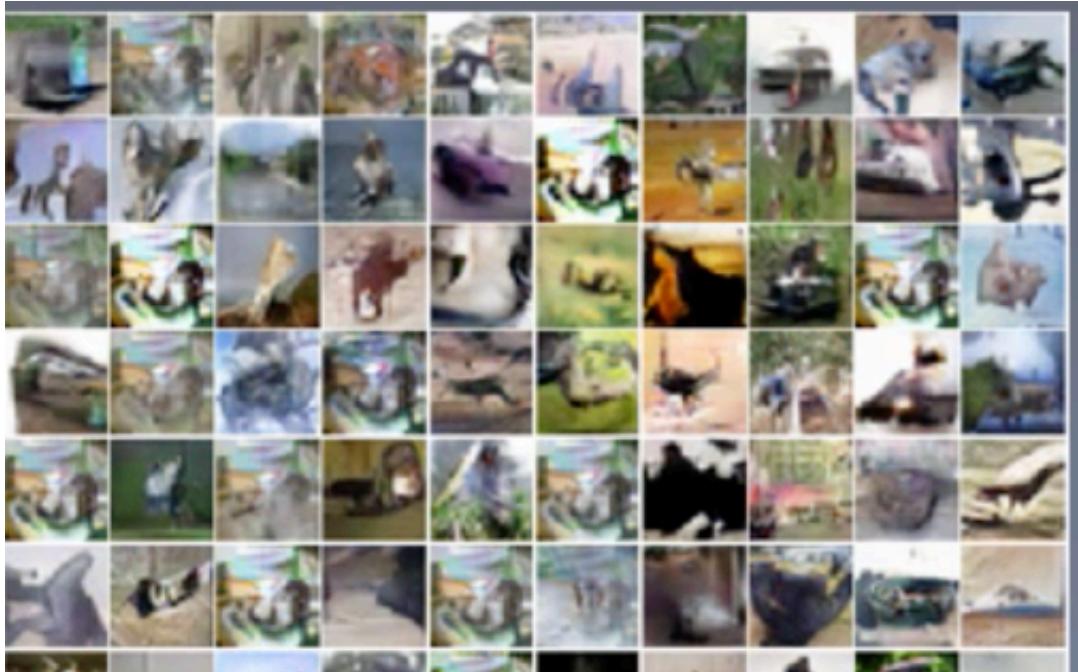
Model	Number of incorrectly predicted test examples for a given number of labeled samples			
	20	50	100	200
DGN [21]		333 ± 14		
Virtual Adversarial [22]		212		
CatGAN [14]		191 ± 10		
Skip Deep Generative Model [23]		132 ± 7		
Ladder network [24]		106 ± 37		
Auxiliary Deep Generative Model [23]		96 ± 2		
Our model	1677 ± 452	221 ± 136	93 ± 6.5	90 ± 4.2
Ensemble of 10 of our models	1134 ± 445	142 ± 96	86 ± 5.6	81 ± 4.3

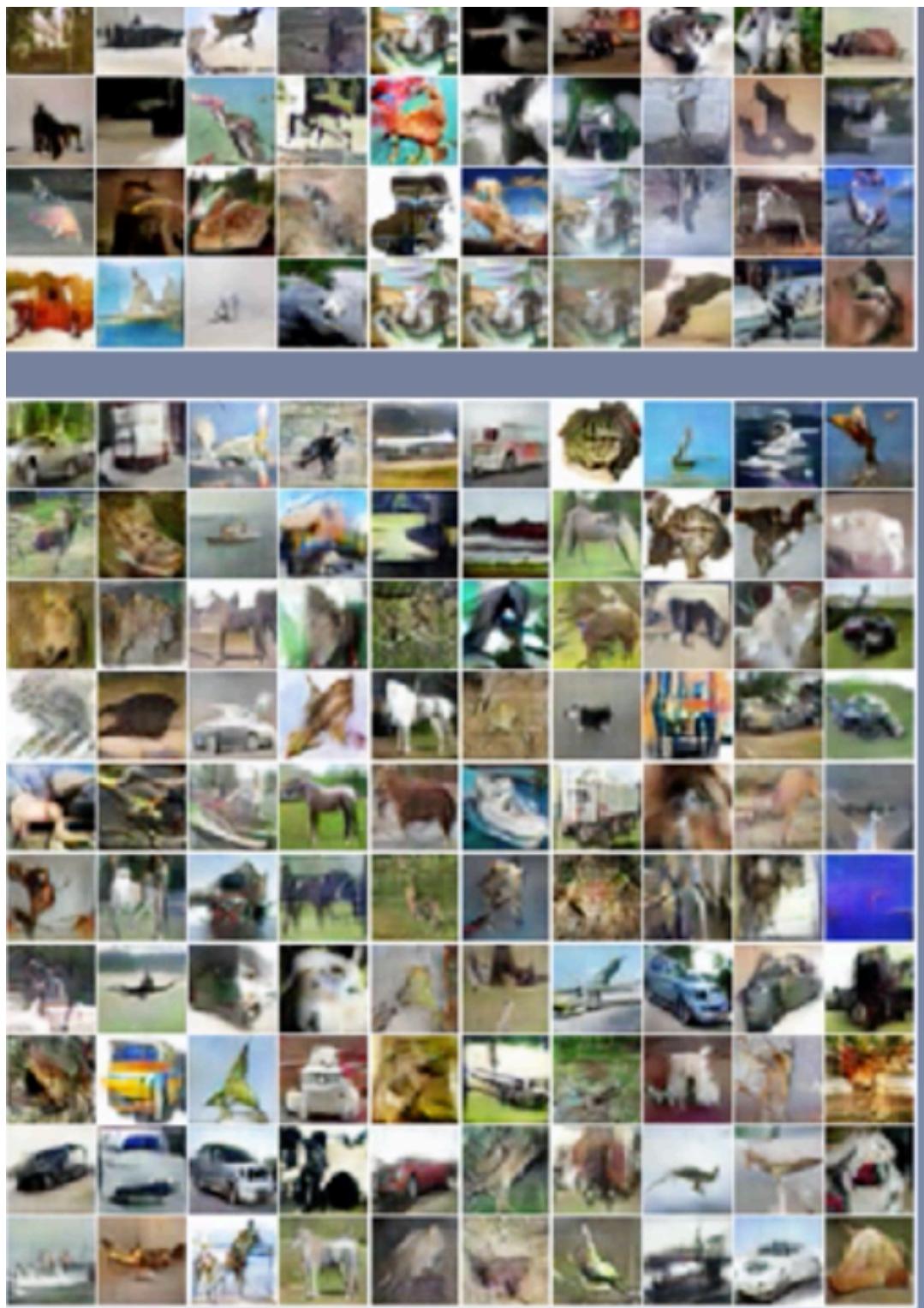


CIFAR-10

- 判别器使用带dropout和权重归一化的9层CNN，生成器使用带BN层的4层CNN
- 使用50%的真实数据和50%由最佳模型生成的数据进行人工评测，众包用户的分类正确率为78.7%，作者们的分类正确率为95%
- 其中Inception score (IS) 得分最高的1%生成图像，众包用户的分类准确率为71.4%

Model	Test error rate for a given number of labeled samples			
	1000	2000	4000	8000
Ladder network [24]			20.40 ± 0.47	
CatGAN [14]			19.58 ± 0.46	
Our model	21.83 ± 2.01	19.61 ± 2.09	18.63 ± 2.32	17.72 ± 1.82
Ensemble of 10 of our models	19.22 ± 0.54	17.25 ± 0.66	15.59 ± 0.47	14.87 ± 0.89





消融实验

来自数据集的真实图像IS得分最高，而所有甚至部分发生模型崩溃的模型IS得分相对较低

下表中的Our methods，应用了除特征匹配和历史平均之外，上文提到的所有优化方法

IS得分只能用来做模型生成图像质量的粗略评估指标，而不能加入到训练中损失函数中，否则将产生对抗性样本，使得IS得分失效

Samples							
Model	Real data	Our methods	-VBN+BN	-L+HA	-LS	-L	-MBF
Score \pm std.	11.24 \pm .12	8.09 \pm .07	7.54 \pm .07	6.86 \pm .06	6.83 \pm .06	4.36 \pm .04	3.87 \pm .03

SVHN

使用与CIFAR-10相同的网络结构和实验设置

Model	Percentage of incorrectly predicted test examples for a given number of labeled samples		
	500	1000	2000
DGN [21]	36.02 \pm 0.10		
Virtual Adversarial [22]		24.63	
Auxiliary Deep Generative Model [23]		22.86	
Skip Deep Generative Model [23]		16.61 \pm 0.24	
Our model	18.44 \pm 4.8	8.11 \pm 1.3	6.16 \pm 0.58
Ensemble of 10 of our models		5.88 \pm 1.0	



ImageNet

- 使用ILSVRC2012数据集的128*128图像，包含1000个类别，此前还没有尝试过
- 由于GAN倾向于低谷分布的熵值，因此大量的类别对于GAN有很大的挑战性
- 在Tensorflow的DCGAN实现基础上进行了大量的修改，并使用多GPU以提高性能
- 原始的DCGAN学习到了一些基本的图像统计信息，生成了具有某种自然颜色和纹理的连续形状，但没有学到任何物体
- 使用文本中描述的技术，能够生成一些看起来像动物，但没有正确结构的图像

128 × 128



DCGAN



ITGAN

结论

生成式对抗性网络是一类很有潜力的生成式模型，但迄今为止一直存在着训练不稳定和缺乏合适评价指标的问题，本文工作提供了部分解决方案

提出了集中提升训练稳定性的技术，使以前无法收敛的模型可以稳定训练

提出的评价指标Inception score可以用来比较模型的生成效果

将这些技术应用与半监督学习中，在多个数据集上取得了state-of-the-art

希望在未来的工作中进行更严格的理论推导

论文总结

A 关键点

- 在那时搁置理论，从直觉出发进行多方面的实验
- 从相关工作中寻找灵感，发现不同图像任务之间的关联性

B 创新点

- 针对batch内的问题进行改进
- 使用图像分类网络进行生成图像质量的评价

C 启发点

- 随着深度学习的快速发展，许多直觉上work的技巧发挥了巨大作用
- 图像理解是图像生成的基础，生成任务是理解任务的自然延伸