软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, [doi: 10.13328/j.cnki.jos.006156]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于一种条件熵距离惩罚的生成式对抗网络研究*

谭宏卫[1,2], 王国栋[1], 周林勇[2], 张自力[1,3]

[1](西南大学 计算机与信息科学学院, 重庆 400715)

[2](贵州财经大学 数统学院, 贵州 贵阳 550025)

[3](School of Information Technology, Deakin University, Locked Bag 20000, Geelong, VIC 3220, Australia)

通讯作者: 张自力, E-mail: zhangzl@swu.edu.cn

摘 要: 生成高质量的样本一直是生成式对抗网络(Generative Adversarial Networks, GANs)领域的主要挑战之一.鉴于此,本文利用条件熵构建一种距离,并将此直接惩罚于 GANs 生成器目标函数,在尽可能地保持熵不变的条下,迫使生成分布逼近目标分布,从而大幅地提高网络生成样本的质量.除此之外,本文还通过优化 GANs 的网络结构以及改变两个网络的初始化策略,来进一步提高 GANs 的训练效率.在多个数据集上的实验结果显示,本文所提出的算法显著提高了 GANs 生成样本的质量;尤其在 CIFAR10, STL10 和 CelebA 数据集上,将最佳的 FID 值从 20.70, 16.15, 4.65 分别降低到 14.02,12.83, 3.22.

关键词: 生成式对抗网络;条件熵距离;网络结构;样本多样性;图像生成

中图法分类号: TP311

中文引用格式: 谭宏卫,王国栋,周林勇,张自力.基于一种条件熵距离惩罚的生成式对抗网络研究.软件学报,2020,31. http://www.jos.org.cn/1000-9825/6156.htm

英文引用格式: Tan HW, Wang GD, Zhou LY, Zhang ZL. Research on generative adversarial networks based on a penalty of a conditional entropy distance. Ruan Jian Xue Bao/Journal of Software, 2020 (in Chinese). http://www.jos.org.cn/1000-9825/6156.htm

Research on Generative Adversarial Networks Based on a Penalty of a Conditional Entropy Distance

TAN Hong-Wei[1,2], WANG Guo-Dong[1], ZHOU Lin-Yong[2], ZHANG Zi-Li[1,3]

[1](School of Computer and Information Sciences, Southwest University, Chongqing 400715, China)

[2](School of Mathematics and Statistics, GuiZhou University of Finance and Economics, Guiyang 550025, China)

[3](School of Information Technology, Deakin University, Locked Bag 20000, Geelong, VIC 3220, Australia)

Abstract: Generating high-quality samples is always one of the most challenges in generative adversarial networks(GANs) field. To this end, in this study, a GANs penalty algorithm is proposed, which leverages a constructed conditional entropy distance to penalize its generator. Under the condition of keeping the entropy invariant, the algorithm makes the generated distribution as close to the target distribution as possible and greatly improves the quality of the generated samples. In addition, to improve the training efficiency of GANs, we optimize the network structure of GANs and change the initialization strategy of the two networks. The experimental results on several datasets show that the penalty algorithm significantly improves the quality of generated samples. Especially, on the CIFAR10, STL10 and CelebA datasets, the best FID value was reduced from 16.19, 14.10, 4.65 to 14.02,12.83, 3.22, respectively.

Key words: generative adversarial networks; conditional entropy distance; network structure; sample diversity; image generation

GANs 是由 Goodfellow 等[1]受博弈论中二人零和博弈思想的启发所提出的一种深度生成模型,其网络结构主要由判别器网络和生成器网络所构成.判别器的目的是尽量正确地判断,输入数据是来自真实数据分布还是

* 基金项目: 国家自然科学基金重点项目,面向需求不确定性的智能服务架构研究(61732019)
收稿时间: 2020-06-13; 修改时间: 2020-07-28; 采用时间: 2020-09-17; jos 在线出版时间: 2020-10-12

来自生成分布;而生成器的目的是尽量去学习与真实数据分布一致的分布,并生成以假乱真的样本.虽然 GANs 已广泛应用于图像生成[2,3]、超分辨率图像合成[4,5]、语义分割[6,7]等多个领域,但要生成高质量的样本仍是该领域的一个挑战.

为了提高 GANs 生成样本的质量,出现了很多针对性的算法[12],这些算法主要基于以下两个角度所提出:

一是基于网络结构的算法研究.由于 GANs 网络结构的灵活性,其判别器网络和生成器网络可为任何类型的网络结构.深度生成式对抗卷积网络(Deep Convolutional Generative Adversarial Networks, DCGAN)[2]就是其中一种典型的网络结构算法,其判别器和生成器均使用卷积神经网络(CNN),同时改变梯度优化算法(Adam)[19],以及加入批量标准化(Batch Normalization, BN)[20]层等策略,来提升 GANs 网络的稳定性及整体性能.此后,研究者们将一些高性能网络模块嵌入到 DCGAN 的网络结构中极大地提升 GANs 网络的性能,如将残差模块[32]整合到 WGAN-GP[14]中来生成文本,自注意机制模块[33]整合到条件 GANs 中来生成图像[17],Laplacian 金字塔模块[5]融入 GANs 网络中来提高人脸图像生成的质量[4]等.除此之外,还有部分算法是针对特定的应用而提出不同结构的 GANs 算法,如 SGAN[21],TripleGAN[22], CycleGAN[23], BigGAN[6]等.这些算法基本上都是从纯网络结构角度来设计,并没有考虑样本多样性问题.本文利用信息熵来衡量样本多样性,并将之融入到算法中来提高 GANs 生成样本的质量.

二是基于目标函数的算法研究.这类算法主要从两个方面研究:(1)是改变目标函数的形式;(2)是惩罚目标函数.针对前者,Nowozin 等[25]将原始 GANs 中的 JS 散度[24]推广到一般化的 f 型散度,并提出 f-GAN 算法;进一步,Mao 等[3]根据原始 GANs 的对抗规则,用平方损失函数来代替 GANs 中的熵损失函数,从而提出 LSGAN 算法.除此之外,Arjovsky 等[26]从微分流形的角度,严格证明 JS 散度是导致 GANs 生成器梯度消失及训练不稳定的主要原因,于是他们提出用 Wasserstein I 型距离[27]代替 JS 散度来衡量真实分布与生成分布之间的距离,并提出 WGAN 算法[28].至于后者,Gulrajani 等[14]利用 1 中心梯度惩罚技术来解决 WGAN 中 Lipschitz 条件的限制,并提出 WGAN-GP 算法来提高 WGAN 生成样本的质量.沿着这条研究路线,Hoang 等[15]开发了零中心梯度惩罚的 GANs 算法(GAN-0GP).Miyato 等[16]提出谱标准化的生成式对抗网络(SNGAN),他们利用标准化的谱范数来限制判别器网络的 Lipschitz 常数,使得网络的 Lipschitz 常数逼近于 1,这相当于对判别器网络实施正则化.截止目前,虽然有很多优秀的 GANs 算法,极大地提升了生成样本的质量,但仍不能满足现实任务的需求,亟需探索高性能的算法来提高 GANs 生成样本的质量.

基于对上述 GANs 算法研究的调查,本文提出一种条件熵距离惩罚的生成式对抗网络,旨在进一步提高 GANs 生成样本的质量;与其他惩罚技术最大的区别在于:罚函数直接惩罚生成器,而非判别器.首先,利用条件熵构造一种距离,可证此距离满足度量空间中的三大条件:正定性,对称性及三角不等式.为了既能保证生成数据多样性与真实数据多样性的一致性,又能迫使生成分布尽可能地逼近真实分布,本文直接用这个距离去惩罚 GANs 生成器.除此之外,本文在 DCGAN[2]的网络结构基础上,进一步优化 GANs 网络结构及初始化策略,主要的优化策略有:(1)将批量标准化(BN)和谱标准化(SN)有机地融入到判别器网络中;(2)删除生成器中的尺度不变层(311 层),即卷积核为 3,步长为 1 以及加边数为 1 的卷积层;(3)改变两个网络的初始化策略,判别器和生成器均使用正交初始化[30].这样的网络结构设计及初始化策略,不但能降低网络的参数空间和减少显存消耗,而且还能提高网络的训练效率和性能.

本文余下内容安排如下:第 1 节,简要介绍 GANs 基础知识;第 2 节,构建一种距离,优化网络结构,并提出一种 GANs 惩罚算法;第 3 节,实验;第 4 节是本文小结.

1 背景知识

GANs 是一种有效而直接的深度生成模型,其基本网络结构是由判别器网络 D 和生成器网络 G 所构成.本质上,GANs 的优化问题是一个极小极大问题[1],其损失函数分为两个部分,分别对应于判别器网络和生成器网络的损失函数,其损失函数分别如下:

$$L_D = -E_{\boldsymbol{x}\sim p_{data}}[\log D(\boldsymbol{x})] - E_{z\sim p_z}[\log(1-D(G(\boldsymbol{z})))], \tag{1}$$

$$L_{G_1} = E_{z\sim p_z}[\log(1-D(G(\boldsymbol{z})))], \tag{2}$$

其中, p_{data} 和 p_z 分别表示真实数据分布和隐分布(先验分布).为便于表示,令 $\hat{\boldsymbol{x}} = G(\boldsymbol{z})$, p_G 表示生成分布, 则有 $\hat{\boldsymbol{x}}\sim p_G$.从理论上讲,GANs 经过多轮迭代,可以使得生成分布无限逼近真实分布[1].但在训练之初,方程(2)中的损失函数有可能达到饱和,无法传递有价值的信息,使得两个分布无法逼近.鉴于此,Goodfellow 等[1]建议将饱和的损失函数(2),转化成非饱和损失函数,于是有

$$L_{G_2} = -E_{z\sim p_z}[\log D(G(\boldsymbol{z}))], \tag{3}$$

相比方程(2),方程(3)中的损失函数更能使 GANs 训练稳定,也因此有研究者将原始 GANs[1]称为非饱和型 GANs(Non-Saturating GANs, NSGAN)[31].

GANs 的训练过程分为两个阶段:第一阶段训练判别器 D,第二阶段训练生成器 G.当训练完判别器 D 后,随后传递真假信息给生成器 G,而生成器 G 根据信息(实质上是梯度信息)的真伪,调整更新策略,尽量生成高质量的样本去“哄骗”判别器.于是,产生了这样的对抗策略:当训练判别器 D 时,尽量使 $D(G(\boldsymbol{z}))=0$;而当训练生成器 G 时,尽量使 $D(G(\boldsymbol{z}))=1$.通过这样的对抗策略, 判别器和生成器都在不断地提升彼此的判别能力和生成能力, 直到判别器无法判断生成器生成的样本是来源于真实数分布还是生成分布.图 1 是 GANs 的训练框架图.

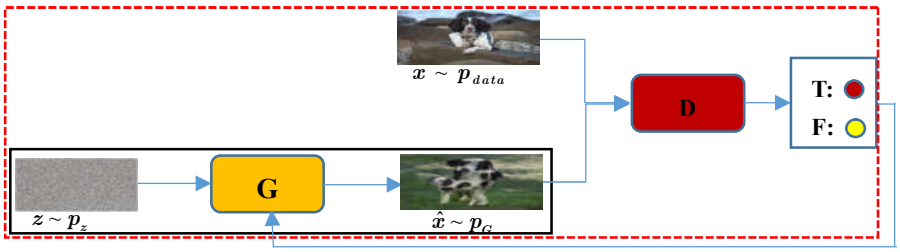


Fig.1 The framework for training GANs: training D in first stage (in the red dotted box) and training G in second stage (in the black solid line box)

图 1 GANs 的训练框架:第一阶段训练判别器 D(红虚线框内), 第二阶段训练生成器 G(黑实线框内)

2 条件熵距离惩罚的生成式对抗网络

在本节内容中,将详细阐述本文所提出的惩罚算法.首先,利用条件熵构建一种距离,并将之直接惩罚于生成器的非饱和损失函数上,即方程(3);其次,在 DCGAN 网络结构[2]的基础上,优化 GANs 的网络结构及超参设置,改变生成器网络和判别器网络的初始化策略,以此来提升模型的训练效率及性能.

2.1 条件熵距离

信息熵是随机变量不确定程度的度量;它也是从平均意义上描述随机变量所需信息量的度量.设 X 是离散型随机变量,其分布函数为 $F_X(x)$,则 X 的信息熵[24]为

$$H(X) = -\sum_{x\in\mathcal{X}} F_X(x)\log F_X(x), \tag{4}$$

其中, $\mathcal{X}$ 表示随机变量 X 的取值空间.同理,对于离散型随机变量 Y ,有 $H(Y) = -\sum_{y\in\mathcal{Y}} F_Y(y)\log F_Y(y)$.其中 $F_Y(y)$ 是随机变量 Y 的分布函数, $\mathcal{Y}$ 是 Y 的取值空间.若对于在 X 给定的条件下,随机变量 Y 的条件熵[24]定义为

$$H(Y\mid X) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} F(x,y)\log F(y\mid x), \tag{5}$$

其中, $F(x,y)$ 和 $F(y|x)$ 分别表示 X , Y 的联合分布函数和条件分布函数.若要度量两个随机变量分布之间的距离,可用它们之间的 KL 散度[24]来度量,其定义如下:

$$D_{KL}(F_X \| \hat{F}_X) = \sum_{x\in\mathcal{X}} F_X(x)\log\frac{F_X(x)}{\hat{F}_X(x)}. \tag{6}$$

基于连续性考虑, 约定 $\log 0/0 = 0$, $\log 0/\hat{F}_X(x) = 0$, $\log F_X(x)/0 = \infty$ .KL 散度 $D(F_X \| \hat{F}_X)$ 度量了真实分布 F_X 与假设分布 $\hat{F}_X$ 之间的距离,可证这个度量是非负的,且当且仅当 $F_X = \hat{F}_X$ 时, $D(F_X \| \hat{F}_X) = 0$.但是,由于 KL 散度不满足对称性和三角不等式,因此它并非真正的距离.在原始 GANs 中,就是用 KL 散度的对称版本 JS 散度来度量真实数据分布和生成分布之间的距离.

现考虑度量两个随机变量 X , Y 之间的关系或一个随机变量包含另一个随机变量的信息量,可用两个随机变量之间的互信息来度量[24],其定义为

$$I(X;Y) = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} F(x,y)\log\frac{F(x,y)}{F_X(x)F_Y(y)}. \tag{7}$$

由 KL 散度(6)定义可知, $I(X;Y) = D(F(x,y) \| F_X(x)F_Y(y))$.由此可见, $I(X;Y)$ 也度量了两个随机变量之间的距离;同样 $I(X;Y)$ 非负,且当且仅当 $X = Y$ 时, $I(X;Y) = 0$.但是, $I(X;Y)$ 同样不是两个随机变量之间真正的距离,因为它不满足距离定义中的三角不等式.鉴于此,根据方程(5)可构造如下距离:

$$\rho(X,Y) = H(X|Y) + H(Y|X). \tag{8}$$

而又由 $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$,可得 $\rho(X,Y) = H(X) + H(Y) - 2I(X,Y)$.由方程(8)可知, $\rho(X,Y)$ 是由条件熵所构成,故称之为条件熵距离(Conditional Entropy Distance, CED).特别指出的是,如果 X,Y 为连续性随机变量,只需将每个定义中的分布函数换成概率密度函数,求和符号换成相应的积分符号即可.

定理 1 $\rho(X,Y)$ 是一种距离,即满足:

(1) 正定性: $\rho(X,Y) \geq 0$,当其仅当 $X = Y$ 时, $\rho(X,Y) = 0$.

(2) 对称性: $\rho(X,Y) = \rho(Y,X)$.

(3) 三角不等式:对于随机变量 X,Y,Z ,有 $\rho(X,Y) + \rho(Y,Z) \geq \rho(X,Z)$.

证明:(1)正定性.由于熵是非负,得 $\rho(X,Y) = H(X|Y) + H(Y|X) \geq 0$,并且当 $X = Y$ 时, $H(X|Y) = H(Y|X) = 0$,可 $\rho(X,Y) = 0$;反之,当 $\rho(X,Y) = 0$ 时,由 $\rho(X,Y) = H(X|Y) + H(Y|X)$ 及熵的非负性,有 $H(X|Y) = H(Y|X) = 0$,此时 $X = Y$.因此, $\rho(X,Y) \geq 0$,当其仅当 $X = Y$ 时, $\rho(X,Y) = 0$.

(2)对称性.由 $\rho(X,Y) = H(X|Y) + H(Y|X) = H(Y|X) + H(X|Y) = \rho(Y,X)$,易知 $\rho(X,Y) = \rho(Y,X)$.

(3)三角不等式.若要证 $\rho(X,Y) + \rho(Y,Z) \geq \rho(X,Z)$,只需证

$$H(X|Y) + H(Y|X) + H(Y|Z) + H(Z|Y) \geq H(X|Z) + H(Z|X),$$

成立.而对于随机变量 X,Y,Z ,有

$$\begin{aligned} H(X|Y) + H(Y|Z) &\geq H(X|Y,Z) + H(Y|Z) \\ &= H(X,Y|Z) \\ &= H(X|Z) + H(Y|X,Z) \\ &\geq H(X|Z). \end{aligned}$$

同理可得, $H(Y|X) + H(Z|Y) \geq H(Z|X)$, 合并两个不等式就有

$$H(X|Y) + H(Y|X) + H(Y|Z) + H(Z|Y) \geq H(X|Z) + H(Z|X),$$

成立,即有 $\rho(X,Y) + \rho(Y,Z) \geq \rho(X,Z)$ 成立.因此, $\rho(X,Y)$ 是一种距离.证毕.

2.2 生成器的条件熵距离惩罚

GANs 生成样本的质量与其样本的多样性和逼真度密切相关.鉴于此,需构建这样一种 GANs 算法:在保持多样性的同时,尽量使得生成分布无限逼近真实分布.利用条件熵距离能实现这样的算法,信息熵可度量样本的多样性,而由此构造的距离可度量生成分布与真实分布之间的距离.为此,只需将条件熵距离直接惩罚于 GANs

算法 1. 条件熵距离惩罚 GANs 算法(EDGAN).

EDGAN 算法利用小批量随机梯度下降法训练网络,同时使用 OAdam 梯度优化算法($\beta_1 = 0.5, \beta_2 = 0.9$),判别器和生成器均使用正交初始化,惩罚因子 λ 设为 1.

输入:

- 样本批量数 $m = 64$;
- 两个网络学习率:判别器 lr_D=0.0001,生成器 lr_G=0.0004;
- 隐分布 $\boldsymbol{z} \sim N_{128}(0, 0.02)$;
- 网络迭代总次数 $k = 100000$;
- 判别器迭代次数 $k_D = 2$,生成器迭代次数 $k_G = 1$.

输出:生成样本 $\hat{\boldsymbol{x}} = G(\boldsymbol{z})$.

for k **do**

  for k_D **do**

  - Sample minibatch of m noise samples $\{\boldsymbol{z}^{(1)}, \boldsymbol{z}^{(2)}, ..., \boldsymbol{z}^{(m)}\}$ from noise prior $N_{128}(0, 0.02)$.
  - Sample minibatch of m examples from $\{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, ..., \boldsymbol{x}^{(m)}\}$ data generating distribution p_{data}.
  - Update the discriminator.

$$L_D = -E_{\boldsymbol{x} \sim p_{data}}[\log D(\boldsymbol{x})] - E_{z \sim p_z}[\log(1 - D(G(\boldsymbol{z})))].$$

  end for

  - Sample minibatch of m noise samples $\{\boldsymbol{z}^{(1)}, \boldsymbol{z}^{(2)}, ..., \boldsymbol{z}^{(m)}\}$ from noise prior $N_{128}(0, 0.02)$.
  - Train and estimate $\rho(X_E, \hat{X}_G)$.
  - Update the generator.

$$L_G = -E_{z \sim p_z}[\log D(G(\boldsymbol{z}))] + \lambda \rho(X_E, \hat{X}_G).$$

end for

生成器的目标函数上,使之生成的样本更具多样性和高逼真度,从而提高 GANs 生成样本的质量.在判别器损失函数不变的条件下,将条件熵距离直接加入到生成器非饱和损失函数中,即方程(3)中,有

$$L_D = -E_{\boldsymbol{x} \sim p_{data}}[\log D(\boldsymbol{x})] - E_{z \sim p_z}[\log(1 - D(G(\boldsymbol{z})))], \tag{9}$$

$$L_G = -E_{z \sim p_z}[\log D(G(\boldsymbol{z}))] + \lambda \rho(X_E, \hat{X}_G). \tag{10}$$

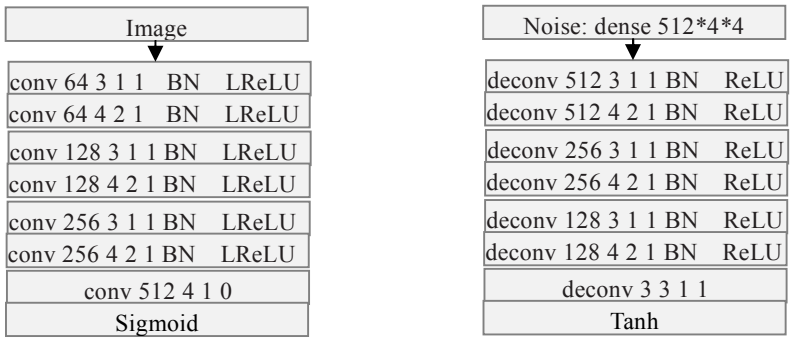
其中, λ 是惩罚因子.在实际中,并不容易估算方程(10)中的条件熵距离,其需要具体的分布函数或密度函数表达式,以图像生成为例,图像数据的真实分布和生成分布并不可知.在这情况下,用经验分布函数代替真实分布函数不失为一个好的选择.设真实数据样本的经验分布函数为 p_{Edata},生成数据样本的经验分布函数为 p_{EG},则方程(10)中的 $X_E \sim p_{Edata}, \hat{X}_G \sim p_{EG}$.由此看出,随机变量 $X_E, \hat{X}_G$ 的取值空间分别是真实数据域和生成数据域.由损失函数(9)和(10)构成的 GANs,我们称之为条件熵距离惩罚的 GANs,简记为熵距离 GANs (Entropy Distance GAN, EDGAN). EDGAN 的算法流程如算法 1 所示.

根据条件熵距离的特性,EDGAN 算法在尽量地保持样本多样性的同时,使得生成分布与真实分布之间的距离尽可能地接近,即在 EDGAN 算法的更新过程中,同时考虑样本的多样性和逼真度两个因素,这有助于提高 GANs 生成样本的质量.虽然,GANs 的目标函数是影响其性能的关键因素,但是绝不是唯一的影响因素,这其中还有网络结构,超参设置及初始化策略等因素都有可能影响其性能.因此,优化 GANs 的网络结构及参数设置也是提升 GANs 性能的重要技术手段之一.本文将在 DCGAN 网络结构[2]的基础上,优化 GANs 的网络结构及超参

设置,改变 GANs 的初始化策略,减小 GANs 的参数空间,加速 GANs 的训练效率,并结合惩罚机制,提升 GANs 的整体性能.

2.3 优化网络结构

DCGAN 的网络结构是一种经典的 GANs 网络结构,其判别器网络和生成器网络均使用卷积神经网络(CNN),它的基本参数设置是:在判别器和生成器中均使用批量标准化(BN),生成器的激活函数除输出层用 Tanh 函数外,其余层都使用 ReLU 激活函数[34],判别器除最后一层用 Sigmoid 激活函数外,其余层均使用 LeaklyRelU 激活函数[35](斜率为 0.2),两个网络的参数初始化都取值于 N(0, 0.02)随机数,两个网络的学习率都是 0.0002,隐分布是均匀分布 U[−1, 1],批量数(Batchsize)是 128,梯度优化算法是 Aadm 算法[19] ($\beta_1 = 0.5, \beta_2 = 0.999$).图 2 是图像尺寸为 3×32×32 的 DCGAN 网络结构,其中在判别器中类似于 conv 64 3 1 1 BN LReLU 的卷积层依次表示该卷积层网络的输出通道是 64(即深度),卷积核(kernel)是 3,步长(stride)是 1,加边数(padding)是 1,卷积操作之后执行批量标准化(BN),最后使用激活函数 LReLU (LeakyLReLU)输出该层网络结果;同理,在生成器中,首先将隐分布随机数(Noise)压缩成尺寸为 512×4×4 的样本,然后传入下一层;deconv 表示降卷积网络层,其他表示的含义与判别器一样.



(a) Discriminator (b) Generator

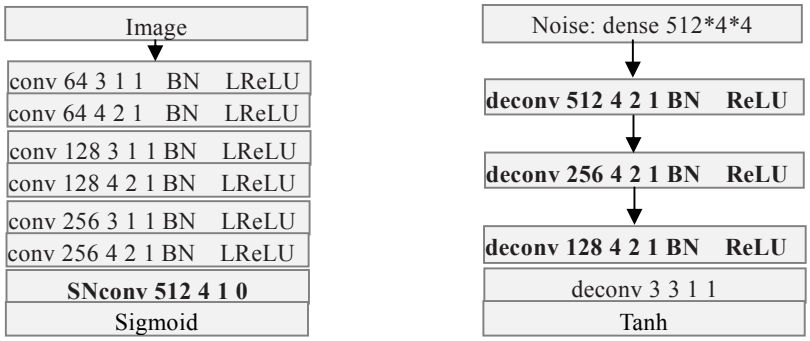
Fig.2 The 3×32×32 network structure of DCGAN: (a)Discriminator, (b)Generator

图 2 图像尺寸 3×32×32 为的 DCGAN 网络结构: (a)判别器, (b)生成器

本文所有的实验都在 Pytorch 框架[36]下完成,为了便于表示和叙述,其网络结构的表示方法也是借助于 Pytorch 中的表示方法.在 Pytorch 中,如果卷积核(kernel)为 3,步长(stride)是 1,加边数(padding)也是 1,则无论是判别器中的卷积层还是生成器中的降卷积层的输出尺寸与上一层的尺寸一致,我们将这种卷积层称为 311 尺度不变层.这种卷积层具有特殊的含义,它只提取特征不作尺度变换.因此,在 DCGAN 的网络结构中,常在 311 尺度不变层后面加上尺度变换层构成另一类输出通道下的样本尺度.这样的设计在每类输出通道下都要进行两次特征提取,使得所提取的特征更加精细化.如果在判别器中,精细化的特征更有利于提升其判别能力,但在生成器中,就不一定能提升生成能力.因为精细化的特征使得生成器更容易忽视样本的一般特征,从而导致生成的样本质量反而下降.后面的实验也证实了这样的判断.

优化网络结构是提升 GANs 性能的技术之一.鉴于上述分析,剔除 DCGAN 生成器中所有的 311 尺度不变层将有助于提升 GANs 生成样本的质量,同时也极大地减小生成器网络的参数空间,加速网络的训练效率.特别强调的是,在图 2 中的生成器的倒数第二层也是 311 层,但其真正目的是通道变换,而非特征提取,所以不能剔除.除此之外,本文还将 SNGAN[16]中的谱标准化(SN)策略融入到 GANs 的判别器中.与 DCGAN 和 SNGAN 最大的不同是,改进的判别器网络并非单纯地使用批量标准化(BN)或谱标准化(SN),而是将两者有机地融合到判别器网络中.具体地说,将 SN 层加入到 Sigmoid 层的前一层,其余层均使用 BN 层.这样设计的目的在于,充分地利用两种标准化的优势.BN 的主要作用是使前后两层之间的分布尽量保持一致,减小每层网络内部之间的协方差偏移(Covariate Shift),改善梯度更新过程,加速网络收敛.而 SN 的主要作用是使得每层网络的 Lipshitz 常数尽量

逼近1,进而使得判别器网络的训练更加稳定,其作用等价于网络正则化.由此看来,BN与SN之间并无冲突,其作用也不可相互替代.同时,无需顾虑SN在网络中的位置,据BN和SN的原理,加入Sigmoid层的前一层是最合理的位置.一般情况下,DCGAN中的Sigmoid层的前一层网络是压缩层,将上一层的结果压缩成N ×1×1×1(N表示通道数),此时使用BN并不能起到任何作用,但SN不一样,仍然能控制该层的Lipschitz常数,此时的SN仍能对模型产生正则化效果.从这个角度分析,SN解决了BN在Sigmoid层的前一层网络中失效的问题.因此,将两者有机融合到判别器中更有利于提升网络的判别能力.图3是优化的DCGAN网络结构.



(a) Discriminator (b) Generator

Fig.3 The optimized network structure of DCGAN(the image size is $3\times32\times32$): (a)Discriminator, (b)Generator

图3 优化的DCGAN网络结构(图像尺寸为$3\times32\times32$): (a)判别器, (b)生成器

初始化策略及超参数设置是GANs网络结构的重要组成部分,这两项配置对网络结果有一定的影响.常用的GANs网络权重初始化是高斯分布(N(0, 0.02)或均匀分布(U(0, 1))随机数.本文采用的初始化策略是判别器与生成器均使用正交初始化[30].这个初始化策略是受到这样的事实所启发:随机向量的正交变换其熵不变[24].特别强调的是,偏置项没有初始化,因为卷积层的后一层加入BN层或SN层之后,该卷积层的偏置项并没有任何作用,所以一般将偏置项去除.除此之外,经过不断的实验,获得如下最优超参:惩罚因子$\lambda=1$,批量数(Batchsize)64,梯度优化算法使用OAdam (Optimistic Adam)算法[52],其参数设为$\beta_1=0.5,\beta_2=0.9$,判别器的学习率lr_D为0.0001,生成器的学习率lr_G为0.0004,隐空间分布是128维的标准高斯分布,网络迭代率是2:1(即生成器每迭代一次判别器需迭代2次).接下来,通过实验来证实,网络结构配置(包括超参数、初始化策略以及结构优化)的高效性,以及惩罚模型的有效性.

3 实验

为了验证EDGAN算法的性能,本文在CIFAR10[41], SVHN[42], STL10[43], CelebA[44], LSUN[45]数据集上进行实验,其中LSUN数据集只用bedroom子集,共3033042张图片,同时与近年来的一些代表性算法作对比.所有实验都在Pytorch框架下完成,且数据集STL10,CelebA和LSUN(bedroom)的图像尺寸被统一剪裁到$3\times32\times32$.首先,在3.1节中简述模型的评价标准;其次,在3.2节中验证优化的网络结构性能;然后,在3.3节中对惩罚模型的有效性进行验证;3.4节是算法性能对比.

3.1 性能评价标准

目前在GANs领域中, IS得分[46]和FID值[47]是两个最经典的性能评价指标,几乎已成为该领域内通用的评价标准.IS得分(Inception Score, IS)是利用在ImageNet数据集上预训练的InceptionV3网络[48]对GANs生成的样本构建一个得分统计量,其表达式如下:

$$IS=\exp(E_{\hat{\boldsymbol{x}}\sim p_G}D_{KL}(p(y\mid\hat{\boldsymbol{x}})\,\|\,p(y)))\,, \tag{11}$$

其中, $D_{KL}(p(y\mid\hat{\boldsymbol{x}})\,\|\,p(y))$ 表示条件类别分布 $p(y\mid\hat{\boldsymbol{x}})$ 与其边缘分布 $p(y)$ 之间的KL散度,可以证明IS得分的自然

对数实质上是类别标签 y 与生成样本 $\hat{\boldsymbol{x}}$ 之间的互信息[49],即有 $\ln(IS)=I(y,\hat{\boldsymbol{x}})$.因此, IS 得分值越大,GANs 模型的性能越好.虽然 IS 得分是最常用的评价指标之一,但进一步研究发现[49],IS 得分并不是最佳的评价标准,甚至当 GANs 生成样本的质量很差时,也有可能获得高得分.

FID(Frechet Inception Distance, FID)值实质上是两个假设的高斯分布之间的 Wasserstein II 型距离[50].具体地说, FID 值是利用 inception V3 网络(其他 CNN 网络也可行)分别将生成样本与真实样本嵌入到一个特征空间中, 同时假设嵌入的样本服从高斯分布,并分别计算嵌入样本的均值 $\boldsymbol{\mu}_G,\boldsymbol{\mu}_d$ 和协方差 $\boldsymbol{C}_G,\boldsymbol{C}_d$,则两个分布之间的 FID 值为

$$FID=\|\boldsymbol{\mu}_G-\boldsymbol{\mu}_d\|_2^2+Tr(\boldsymbol{C}_G+\boldsymbol{C}_d-2(\boldsymbol{C}_d\boldsymbol{C}_G)^{1/2}). \tag{12}$$

由 FID 值的定义(12)可知,只要满足高斯分布的假设,FID 值能很好地度量 GANs 模型的性能.根据中心极限定理[51],若样本容量趋于无穷大,则其极限分布服从高斯分布.于是,在实际中计算 FID 时常要求 GANs 生成足够多的样本.由于 FID 值是间接地度量生成分布与真实分布之间的距离, 因此 FID 越小,模型的性能越好.相比 IS 得分,FID 值更能全面地度量 GANs 模型生成样本的质量(无论是多样性还是逼真度)[47],这个指标几乎已成为 GANs 领域内统一的评价标准.**本文所有的实验结果都仅以 FID 值作为评价标准.**在没有特别说明的情况下,所有实验的 FID 值,都是分别从真实样本中抽取 50000 个样本,从生成样本中抽取 50000 个样本来计算.

3.2 优化的网络结构性能验证

为了验证两种优化策略的有效性,即(1)将判别器倒数第二层的卷积层修改为谱标准化(SN)的卷积层(如图 3(a)倒数第二层);(2)删除生成器中的 311 不变层,在 DCGAN 网络结构上执行消融实验(Ablation Study).为便于表示,我们用 SN_in_P(SN in Penultimate)和 R311(Remove 311)分别表示第(1)和(2)种优化策略.为简化实验过程,这个消融实验仅在 CIFAR10 数据集上执行,生成器分别训练 70k,100k,140k 和 200k 次.然后,将这种优化策略拓展到其他 4 个数据集上来验证其性能. CIFAR10 数据集是由 60000 张 32×32 彩色图片所构成,其中有 50000 张图片构成训练集,10000 张图片构成验证集,本文用到的数据集是去标签后的训练集(50000 张).

Table 1 Testing the effectiveness of optimization strategy on network structure(FID)

表 1 网络结构优化策略的有效性验证(FID)

项 目	70k	100k	140k	200k
DCGAN	121.54	113.12	110.56	109.21
SN_in_P	**91.54**	**101.43**	**98.12**	**99.23**
R311	**30.60**	**27.36**	**26.27**	**32.06**
SN_in_P+R311(128)	**26.35**	**24.23**	**22.59**	**26.55**
SN_in_P+R311(64)	**20.09**	**19.58**	**17.70**	**20.13**

表 1 是两种优化策略的消融实验结果.表 1 的结果显示,SN_in_P 策略对 DCGAN(表 1 第二行)的整体性能有所提升,但是提升幅度显著差于 R311 策略,尤其是训练次数为 140k 时,FID 值下降至 26.27.可以看出,移除前(表 1 DCGAN)和移除后(表 1 R311)的结果形成了鲜明的对比.为了验证两种策略相结合的效果,我们分别实验批量数为 128(表 1 第五行)和 64(表 1 第六行)时,网络的性能变化情况.可以看出,在四种迭代次数下,两种优化策略相结合的网络性能均有不同程度的提升,FID 值均在下降.相比之下,批量数为 64 时,网络效果更好.表 1 的结果充分地证实,本文所提出的两种优化策略是有效的,尤其是两种优化策略相结合的网络更是大幅地提升了网络的性能.

为全面地验证两种优化策略的效果,我们分别在 CIFAR10、SVHN、CelebA 等 5 个数据集上训练两种策略相结合的网络.为便于表示,称这个优化的网络结构为 OptimizedDCGAN.在这个实验中,将批量数(Batchsize)设为 64.表 1 的结果已经显示,批量数为 64 比批量数为 128 的网络的性能较好.为验证这个事实,单独对批量数执行消融实验.图 4 是 DCGAN 分别在 CIFAR10 数据集上训练 70k、100k、140k 和 200k,批量数分别为 64 和 128 的结果.可以清晰地看出,批量数为 64 的网络(蓝色)显著优于批量数为 128 的网络(浅灰色).没有特别说明的情

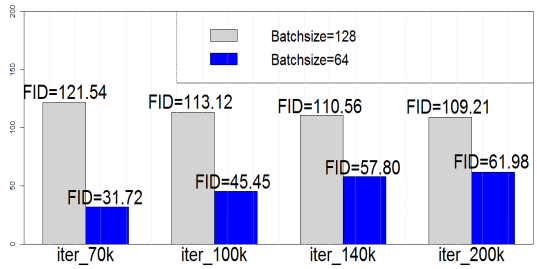


Fig.4 The effect of batchsize on network performance(CIFAR10)

图 4 批量数对网络性能的影响(CIFAR10)

Table 2 The performance of the Optimized DCGAN network structure (FID value)

表 2 优化的 DCGAN 网络结构的性能(FID 值)

数据集	网络结构	70k	100k	140k	200k
CIFAR10	DCGAN	121.54	113.12	110.56	109.21
	OptimizedDCGAN	**20.09**	**19.58**	**17.70**	**20.13**
SVHN	DCGAN	6.42	11.19	71.07	41.72
	OptimizedDCGAN	**5.10**	**4.72**	**4.71**	**4.85**
STL10	DCGAN	56.79	55.74	53.98	58.98
	OptimizedDCGAN	**16.57**	**16.39**	**16.44**	**15.28**
CelebA	DCGAN	16.71	11.23	7.54	8.19
	OptimizedDCGAN	**6.34**	**5.25**	**4.89**	**4.54**
LSUN	DCGAN	28.33	17.53	15.12	18.11
	OptimizedDCGAN	**10.12**	**8.77**	**7.12**	**7.50**

Table 3 Different penalty factors and corresponding FID values on CIFAR10

表 3 不同惩罚因子 λ 所对应的 FID 值(CIFRA10)

惩罚因子 λ	0.1	1	5	10	15	20
FID 值	16.18	**14.02**	16.59	17.01	15.86	17.53

况下,后续所有实验的批量数均设为 64.表 2 是 OptimizedDCGAN 在 5 个数据集上分别训练 70k,100k,140k 和 200k 次后的结果.相比 DCGAN 网络结构,优化的网络结构 OptimizedDCGAN 的性能在 5 个数据集上均有显著地提升.值得强调的是,在 SVHN 数据上,DCGAN 在四种训练次数上表现极不稳定,尤其是训练次数为 140k 时,FID 值忽然从 11.19(100k)上升至 71.07,200k 时又下降至 41.72,而 OptimizedDCGAN 网络结构表现较为稳定.表 2 的结果再次验证了本文所提出的网络结构优化策略的有效性.

3.3 惩罚模型的有效性验证

在这个部分内容，验证由 OptimizedDCGAN 网络结构与条件熵惩罚相结合的惩罚模型的有效性.首先,探究惩罚模型一些关键参数的最优配置,包括迭代率、学习率以及初始化策略等;其次,验证本文所提出惩罚技术的有效性.

3.3.1 惩罚模型的参数配置及初始化策略

考虑到深度神经网络对网络参数配置的敏感性,我们通过一系列的消融实验来确定一些关键的超参数及其初始化策略.通过反复实验得到,惩罚模型的一些最优参数配置如下:惩罚因子 $\lambda=1$,迭代率 2:1(即生成器每更新一次判别器需更新两次),判别器学习率 lr_D=0.0001,生成器学习率 lr_G=0.0004,梯度优化算法使用 OAdam(0.5,0.9)，初始化策略:判别器和生成器均使用正交初始化.下面,将利用消融策略逐一验证这些配置的有效性.为简化验证过程,所有消融实验仅在 CIFAR10 上执行,生成器均训练 100k 次.

首先,确定最优的惩罚因子 λ.在其他参数保持不变的前提下,观察 $\lambda=0.1$，1，5，10，15，20 时网络性能变化情况,其结果显示在表 3 中.可以看出, 当 $\lambda=1$ 时,网络性能表现最佳,此时 FID 值为 14.02.其次,再观察网络学习

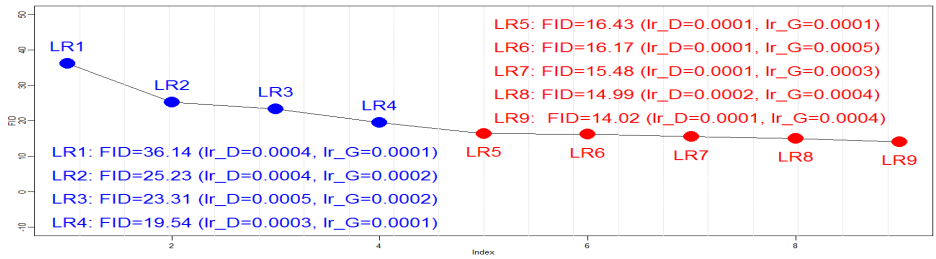


Fig.5 The effect on network performance with different learning rates

图 5 不同学习率对网络性能的影响

Table 4 Implementing the ablation studies with respect to some hyper parameters and initialization (CIFAR10)

表 4 一些超参数及初始化策略的消融实验(CIFRA10)

项 目	A	B	C	D	E
学习率	**14.02** **(0.0001, 0.0004)**	36.14 (0.0004, 0.0001)	15.48 (0.0001, 0.0003)	25.23 (0.0004, 0.0002)	16.02 (0.0002, 0.0002)
迭代率	16.75 (1:1)	**14.02** **(2:1)**	16.00 (3:1)	16.53 (4:1)	17.52 (5:1)
初始化	**14.02** **(Orth, Orth)**	18.20 (Orth, Glorot)	22.52 (N(0.02),N(0.02))	17.31 (KaiN, Orth)	17.91 (N(0,0.02), Orth)
优化算法	**14.02** **OAdam(0.5, 0.9)**	15.26 Adam(0.5, 0.9)	16.26 OAdam(0.5,0.999)	18.55 Adam(0.5,0.999)	17.60 OAdam(0.9, 0.9)

率、迭代率、初始化策略、梯度优化算法对网络性能的影响,每个目标观察量均设置 5 种变化,并统一表示为 A、B、C、D、E.这些实验结果均总结在表 4 中,其中每个目标观察量的 5 种变化 A-E 的含义均置于 FID 值的下方.由表 4 的结果可知,当 lr_D=0.0001,lr_G=0.0004 时,网络性能表现最佳.从学习率的实验结果,还可看出,这个参数对网络性能的影响较大;特别地,当 lr_D>lr_G 时(如表 4 中 B,D 组合),网络性能较差,反之,则较好(如 A、C、E 组合).为此,我们单独执行实验 9 种学习率组合探究这个有趣的现象,结果如图 5 所示.在图 5 中,蓝色的点(LR1-LR4)表示 lr_D > lr_G 的点,红色的点表示 lr_D ≤ lr_G 的点.可以清晰地看出,蓝色点的 FID 值普遍高于红色的点,这说明上述现象是存在的.对于迭代率,选择 2:1 较好.在初始化方法中,选用 4 种初始化方法:正交初始化(Orth)[30]、Glorot 正态初始化(Glorot)[53]、高斯初始化 N(0,0.02)以及 Kaiming 正态初始化(KaiN)[55],5 种组合.表 4 的结果显示,当判别器和生成器均实施正交初始化时,网络性能最优.在梯度优化算法中,只选择两种优化算法(Adam, OAdam),不同参数下的 5 种组合.已有研究表明,Adam 算法一般要优于其他优化算法[14,15,54],如 SGD,RMSprop,Rprop 等算法.在惩罚模型中,选择 OAdam(0.5,0.9)较优.

3.3.2 惩罚技术的有效性验证

为了充分地体现本文所提出惩罚技术的有效性,分别在 CIFAR10、SVHN、STL10、CelebA 和 LSUN 五个数据集上执行如下实验:(1)不带任何惩罚的网络(No-Penalty);(2)利用条件熵距离惩罚生成器的网络(EDGAN);(3)利用 Jensen-Shannon(JS)散度(对称版本的 KL 散度)惩罚生成器的网络(With-JS-Penalty);(4)利用 WGAN-GP[14]中的梯度惩罚技术惩罚判别器,条件熵距离惩罚生成器,即双惩罚网络(With-Bi-Penalty),其中判别器惩罚因子沿用 WGAN-GP 中的设置(惩罚因子为 10, Lipschitz 常数为 1).同样,所有生成器均训练 100k 次,实验结果如表 5 所示.特别强调的是,表 5 第二列的结果(不带惩罚的网络)与表 2 训练 100k 次时的结果不一样,主要是由于两个网络使用的参数及初始化策略不一致,前者使用 3.3.1 的参数及初始化配置,而后者使用 DCGAN 中的配置. 首先,观察不带惩罚的网络(表 5 第二列)与带条件熵距离惩罚生成器的网络(表 5 第三列,本文提出的算法),5 个数据集所对应的 FID 值均有不同程度的下降,网络性能获得提升.这充分地说明,本文所提出的惩罚技术是有效的.其次,再观察带 JS 惩罚的网络(表 5 第四列)以及双惩罚网络(表 5 最后一列),5 个数据集上的 FID 值均大于第三列的 FID 值(EDGAN),这充分地证实,本文所提出的惩罚技术具有明显的优势.

Table 5 Verifying the penalty effect of EDGAN algorithm: the generator iterates 100k times (FID value)

表 5 验证 EDGAN 算法的惩罚效果: 生成器更新 100k 次的 FID 值

Datasets	No-Penalty	EDGAN(Ours)	With-JS-Penalty	With-Bi-Penalty
CIFAR10	17.92	**14.02**	31.63	27.84
SVHN	4.24	**3.88**	10.94	7.10
STL10	15.03	**12.83**	35.22	31.58
CelebA	4.41	**3.22**	12.02	7.39
LSUN	5.93	**5.13**	10.45	8.98

Table 6 Comparison of the algorithm performance

表 6 算法性能对比

算法	CIFAR10	SVHN	STL10	CelebA	LSUN	iteration times
LSGAN (2017)	21.64	3.50	19.17	4.70	**4.65**	200k
WGAN-GP (2017)	21.89	4.52	16.15	4.56	12.65	200k
SNGAN (2018)	20.70	6.53	40.15	7.56	26.98	100k
SAGAN (2018)	35.18	26.91	29.77	9.49	42.26	200k
GAN-0GP (2019)	16.19	**3.07**	14.10	4.93	6.10	100k
EDGAN (Ours)	**14.02**	3.88	**12.83**	**3.22**	5.13	100k
Real Datasets	0.45	0.24	0.84	0.34	0.55	

3.4 整体性能对比

通过上述一系列实验已验证,EDGAN 算法能显著提高 GANs 网络的性能.为了更进一步验证 EDGAN 的优越性,现将该算法与近年来的一些代表性算法作比较.本文在 5 个数据集上训练 LSGAN(2017),WGAN GP(2017)和 SNGAN(2018)等 5 个算法,其结果显示在表 6 中.其中,在表 6 中的最后一列表示最优结果的生成器迭代次数,最后一行是真实数据的 FID 值.从表 6 中的结果看出,EDGAN 算法在多个数据集上都超越了目前的一些经典算法,只有在 SVHN 和 LSUN 数据集上稍逊于 GAN-0GP 和 LSGAN 算法.与此同时, EDGAN 在 CIFAR10, STL10 和 CelebA 数据集上的 FID 值更进一步地接近真实数据集的 FID 值.在表 6 中加灰色的行是 SAGAN 算法,这是一个经典的条件 GANs 算法,而这里是将其标签信息从算法中删除后,此时 SAGAN 就退化为无条件的 GANs(即 GANs).在此执行这个实验,意在说明是否可以将条件 GANs 的优良性能移植到无条件 GANs 上? SAGAN 在 5 个数据集上实验结果显示,答案是否定的.这也充分地证实这样的一个事实:条件 GANs 并不能当作无条件 GANs 使用.因此,开发高性能的无监督图像生成算法,正是本文研究的出发点.

4 结束语

本文利用条件熵距离对 GANs 生成器的目标函数进行惩罚,旨在保持样本多样性的条件下迫使生成分布尽可能地接近真实分布,这使得 GANs 生成的样本既有多样性又具高逼真度,从而提高 GANs 生成样本的整体质量.除此之外,本文在 DCGAN 网络结构的基础上,对 GANs 的网络结构进行优化,这其中包括:(1)根据批量标准化(BN)和谱标准化(SN)的互补特性,将两个标准化技术有机地融入判别器网络中,来提高网络的判别能力;(2)去除生成器网络中的 311 尺度不变层,提升生成器的生成能力,同时也能减小生成器网络的参数空间,以此提高网络的训练效率及性能;(3)改变两个网络的初始化策略,判别器网络和生成器网络均使用正交初始化.最后,将惩罚的 GANs 目标函数与优化的网络结构相结合,形成条件熵距离惩罚的 GANs,即 EDGAN. 据实验结果显示, EDGAN 算法已经超越了目前的一些经典 GANs 算法.

虽然本文通过实验验证 EDGAN 能提高 GANs 生成样本的质量,但其收敛性理论问题仍有待进一步的研究.再者,如何利用 EDGAN 生成的样本去解决一些无监督问题,即 EDGAN 的下游问题,仍有待深入研究.

References:

[1] Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial nets. In: Ghahramani Z, eds. Advances in Neural Information Processing Systems 27. New York: Curran Associates Inc., 2014. 2672-2680.

[2] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. In: Proc. of the 4th Int'l Conf. on Learning Representations(ICLR), 2016. https://openreview.net/group?id=ICLR.cc/2016

[3] Mao XD, Li Q, Xie HR, Lau RK, Wang Z, Smolley SP. Least squares generative adversarial networks. In: Katsushi I, ed. 2017 IEEE International Conference on Computer Vision. Piscataway: IEEE Computer Society, 2017. 2794-2802.

[4] Karras T, Aila T, Laine S, Lehtinen J. Progressive growing of GANs for improved quality, stability, and variation. In: Proc. of the 6th Int'l Conf. on Learning Representations(ICLR), 2018. https://openreview.net/group?id=ICLR.cc/2018

[5] Lai WS, Huang JB, Ahuja N, Yang MH. Deep laplacian pyramid networks for fast and accurate super-resolution. In: Conner LO, ed. 30th IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Computer Society, 2017. 624-632.

[6] Brock A, Donahua J, Simonyan K. Large scale GAN training for high fidelity natural image synthesis. In: Proc. of the 7th Int'l Conf. on Learning Representations (ICLR), 2019. https://openreview.net/group?id=ICLR.cc/2019

[7] Samson L, Noord NV, Booij O, Hofmann M, Gavves E, Ghafoorian M. I bet you are wrong: gambling adversarial networks for structured semantic segmentation. In: Peter W, eds. 2019 IEEE International Conference on Computer Vision. Piscataway: IEEE Computer Society, 2019. 3700-3708.

[8] Lee MCH, Petersen K, Pawlowski N, Glocker B, Schaap M. Tetris: template transformer networks for image segmentation with shape priors. IEEE Trans. on Medical Imaging, 2019, 38(11): 2596 – 2606.

[9] Kumar A, Sattigeri P, Fletcher PT. Semi-supervised learning with gans: Manifold invariance with improved inference. In: Guyon I, eds. Advances in Neural Information Processing Systems 30. New York: Curran Associates Inc., 2017. 5534-5544.

[10] Mukherjee S, Asnani H, Lin E, Kanan S. ClusterGAN : latent space clustering in generative adversarial networks. In: Avinash L, eds. AAAI-19/IAAI-19/EAAI-19 Proceedings. California: AAAI Press, 2019. 4610-4617.

[11] Tolstikhin I, Gelly S, Bousquet O, Simon-Gabriel CJ, Scholkoph B. AdaGAN: boosting generative models. In: Guyon I, eds. Advances in Neural Information Processing Systems 30. New York: Curran Associates Inc., 2017. 5424-5433.

[12] Cao YJ, Jia LL, Chen YX, Lin N, Yang C, Zhang B, Liu Z, Li XX, Dai HH. Recent advances of generative adversarial networks in computer vision. IEEE Access, 2018, 7: 14985-15006. DOI: 10.1109/ACCESS.2018.2886814

[13] Wang B, Liu K, Zhao J. Conditional generative adversarial networks for commonsense machine comprehension. In: Sierra C, ed. Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence. California: IJCAI Information Sciences Institute, 2017. 4123-4129.

[14] Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville A. Improved training of Wasserstein GANs. In: Guyon I, eds. Advances in Neural Information Processing Systems 30. New York: Curran Associates Inc., 2017. 5767--5777.

[15] Thanh-Tung H, Tran T, Venkatesh S. Improving generalization and stability of generative adversarial networks. In: Proc. of the 7th Int'l Conf. on Learning Representations (ICLR), 2019. https://openreview.net/group?id=ICLR.cc/2019

[16] Miyato T, Kataoka T, Koyama M, Yoshida Y. Spectral normalization for generative adversarial networks. In: Proc. of the 6th Int'l Conf. on Learning Representations (ICLR), 2018. https://openreview.net/group?id=ICLR.cc/2018

[17] Zhang H, Goodfellow I, Metaxas D, Odena A. Self-attention generative adversarial networks. In: Chaudhuri K, eds. Proceedings of Machine Learning Research(PMLR). Online Publication, 2019, 97: 7354-7363. http://proceedings.mlr.press/

[18] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang ZH, Karpathy A, Khosla A, Bernstein M, Berg AC, Li FF. ImageNet large scale visual recognition challenge. Int. J. Comput. Vis., 2015, 115(3):211-252

[19] Kingma DP, Ba JL. Adam: a method for stochastic optimization. In: Proc. of the 3rd Int'l Conf. on Learning Representations (ICLR), 2015. https://openreview.net/group?id=ICLR.cc/2015

[20] Ioffe S, Szegedy C. Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Bach F, eds. Proceedings of Machine Learning Research(PMLR). Online Publication, 2015, 37: 448-456. http://proceedings.mlr.press/

[21] Huang X, Li YX, Poursaeed O, Hopcroft J, Belongie S. Stacked generative adversarial networks. In: Conner LO, ed. 30th IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Computer Society, 2017. 5077-5086.

[22] Li CX, Xu T, Zhu J, Zhang B. Triple generative adversarial nets. In: Guyon I, eds. Advances in Neural Information Processing Systems 30. New York: Curran Associates Inc., 2017. 4088--4098.

[23] Zhu JY, Park T, Isola P, Efros AA. Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Katsushi I, ed. 2017 IEEE International Conference on Computer Vision. Piscataway: IEEE Computer Society, 2017. 2223-2232.

[24] Gray RM. Entropy and Information Theory. New York: Springer-verlag, 2013. 17-44.

[25] Nowozin S, Cseke B, Tomioka R. f-GAN: training generative neural samplers using variational divergence minimization. In: Lee DD, eds. Advances in Neural Information Processing Systems 29. New York: Curran Associates Inc., 2016. 271-279.

[26] Arjovsky M, Bottou L. Towards principled methods for training generative adversarial networks. In: Proc. of the 5th Int'l Conf. on Learning Representations (ICLR), 2017. https://openreview.net/group?id=ICLR.cc/2017

[27] Villani C. Optimal Transport: Old and New. Berlin: Springer-verlag, 2008. 107-125.

[28] Arjovsky M, Chintala S, Bottou L. Wasserstein generative adversarial networks. In: Precup D, eds. Proceedings of Machine Learning Research(PMLR). Online Publication, 2017, 70: 214-223. http://proceedings.mlr.press/

[29] Dudley RM. Real Analysis and Probability. Cambridge: Cambridge University Press, 2004. 188-220.

[30] Saxe AM, McClelland JL, Ganguli S. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In: Proc. of the 2nd Int'l Conf. on Learning Representations(ICLR), 2014. https://openreview.net/group?id=ICLR.cc/2014

[31] Fedus W, Rosca M, Lakshminarayanan B, Dai AM, Mohamed S, Goodfellow I. Many paths to equilibrium: GANs do not need to decrease a divergence at every step. In: Proc. of the 6th Int'l Conf. on Learning Representations (ICLR), 2018. htttps://openreview.net/group?id=ICLR.cc/2018

[32] He KM, Zhang XY, Ren SQ, Sun J. Deep residual learning for image recognition. In: Conner LO, ed. 29th IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Computer Society, 2016. 770-778.

[33] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, Kaiser L, Polosukhin I. Attention is all you need. In: Guyon I, eds. Advances in Neural Information Processing Systems 30. New York: Curran Associates Inc., 2017. 5998--6008.

[34] Chapados N, Bengio Y, Vincent P, Ghosn J, Dugas C, Takeuchi I, Meng LY. Estimating car insurance premia: a case study in high-dimensional data inference. In: Dietterich TG, eds. Advances in Neural Information Processing Systems 14. Massachusetts: MIT Press, 2002. 1369-1376.

[35] Maas AL, Hannun AY, and Ng AY. Rectifier nonlinearities improve neural network acoustic models. In: Dasgupta S, eds. Proceedings of Machine Learning Research(PMLR). Online Publication, 2013, 28: 106-114. http://proceedings.mlr.press/

[36] Paszke A, Gross S, Francisco M, Adam L, James B, Gregory C, Trevor K, Lin ZM, Natalia G, Luca A, Alban D, Andreas K, Edward Y, Zachary D, Martin R, Alykhan T, Sasank C, Benoit S, Lu F, Bai, JJ, Chintala S. PyTorch: an imperative style, high-performance deep learning library. In: Wallach H, eds. Advances in Neural Information Processing Systems 32. New York: Curran Associates Inc., 2019. 8026—8037.

[37] Salakhutdinov R, Hinton G. Deep Boltzmann machines. J. Mach. Learn. Res, 2009, 5(2):1967-2006.

[38] Bengio Y, Thibodeau-Laufer E, Alain G, Yosinski J. Deep generative stochastic networks trainable by backprop. In: Xing EP, eds. Proceedings of Machine Learning Research(PMLR). Online Publication, 2014, 32: 214-223. http://proceedings.mlr.press/

[39] Kingma DP, Welling M. Auto-encoding variational Bayes. In: Proc. of the 2nd Int'l Conf. on Learning Representations (ICLR), 2014. https://openreview.net/group?id=ICLR.cc/2014

[40] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature, 1986, 323(6088): 533-536.

[41] Krizhevsky A. Learning multiple layers of features from tiny images [MSc. Thesis]. Toronto, Canada: University of Toronto, 2009.

[42] Netzer Y, wang TJ, Coates A, Bissacco A, Wu BL, Ng AY. Reading digits in natural images with unsupervised feature learning. In: Proc. of the 25th Conf. on Neural Information Processing Systems Workshop on Deep Learning and Unsupervised Feature Learning, 2011.

[43] Coates A, Lee H, Ng AY. An analysis of single layer networks in unsupervised feature learning. In: Gordon G, eds. Proceedings of Machine Learning Research(PMLR). Online Publication, 2011, 15: 215-223. http://proceedings.mlr.press/

[44] Liu ZW, Luo P, Wang XG, Tang XO. Deep learning face attributes in the wild. In: Ruzena B, eds. 2015 IEEE International Conference on Computer Vision. Piscataway: IEEE Computer Society, 2015. 3730-3738.

[45] Kramberger T, Potočnik B. LSUN-stanford car dataset: enhancing large-scale car image datasets using deep learning for usage in GAN training. Appl. Sci., 2020, 10(14): 4913.

[46] Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X. Improved techniques for training GANs. In: Lee DD, eds. Advances in Neural Information Processing Systems 29. New York: Curran Associates Inc., 2016. 2234-2242.

[47] Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. GANs trained by a two time-scale update rule converge to a local nash equilibrium. In: Guyon I, eds. Advances in Neural Information Processing Systems 30. New York: Curran Associates Inc., 2017. 6626--6637.

[48] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Conner LO, ed. 29th IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Computer Society, 2016. 2818-2826.

[49] Barratt S, Sharma R. A note on the inception score. In: Proc. of the Int'l Conf. on Machine Learning Workshop on Theoretical Foundations and Applications of Deep Generative Models, 2018.

[50] Vaserstein LN. Markov processes over denumerable products of spaces describing large systems of automata. Probl.Inform.Trans, 1969, 5(3): 64-72.

[51] Vaart A. Asymptotic Statistics. Cambridge: Cambridge University Press, 1998. 68-101.

[52] Daskalakis C, Iiyas A, Syrgkanis V, Zeng HY. Training GAN with optimism. In: Proceedings of the 6th International Conference on Learning Representations (ICLR), 2018. https://openreview.net/group?id=ICLR.cc/2018

[53] Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In: Teh WY, eds. Proceedings of Machine Learning Research(PMLR). Online Publication, 2010, 9: 249-256. http://proceedings.mlr.press/

[54] Berard H, Gidel G, Almahairi A, Vincent P, Lacoste-Julien S. A closer look at the optimization landscapes of generative adversarial networks. In: Proceedings of the 8th International Conference on Learning Representations (ICLR), 2020. https://openreview.net/group?id=ICLR.cc/2020

[55] He KM, Zhang XY, Ren SQ, Sun J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In: Ruzena B, eds. 2015 IEEE International Conference on Computer Vision. Piscataway: IEEE Computer Society, 2015. 1026-1034.