GP-WIRGAN：梯度惩罚优化的 Wasserstein 图像循环生成对抗网络模型

冯 永[1),2)] 张春平[1),2)] 强保华[3),4)] 张逸扬[1),2)] 尚家兴[1),2)]

[1)](重庆大学计算机学院 重庆 400044)

[2)](重庆大学信息物理社会可信服务计算教育部重点实验室 重庆 400044)

[3)](桂林电子科技大学广西可信软件重点实验室 广西 桂林 541004)

[4)](桂林电子科技大学广西光电信息处理重点实验室培育基地 广西 桂林 541004)

摘 要 通常情形下，现有的图像生成模型都采用单次前向传播的方式生成图像，但实际中，画家通常是反复修改后才完成一幅画作的；生成对抗模型(Generative Adversarial Networks，GAN)能生成图像，但却很难训练. 在保证生成图像质量的前提下，效仿作画时的不断更新迭代，以提升生成样本多样性并增强样本语义，同时引入 Wasserstein 距离，提出了 Wasserstein 图像循环生成对抗网络模型，简称 WIRGAN(Wasserstein Image Recurrent Generative Adversarial Networks Model). WIRGAN 定义了生成模型和判别模型，其中，生成模型是由一系列结构相同的神经网络模型组成的循环结构，用时间步骤 T 控制生成模型的循环次数，用于迭代式生成图像，并以最后一个循环结构的生成图像作为整个生成模型的输出；判别模型也由神经网络构建，结合权重剪枝技术，用来判别输入图像是生成的还是真实的. WIRGAN 利用 Wasserstein 距离作为目标函数，将生成模型和判别模型进行博弈对抗训练. 另外，由于模型存在难以优化的问题，本文引入了梯度惩罚来解决此类问题，进一步提出了梯度惩罚优化的 Wasserstein 图像循环生成对抗网络模型(Gradient Penalty Optimized Wasserstein Image Recurrent Generative Adversarial Networks Model，GP-WIRGAN). 最后，WIRGAN 和 GP-WIRGAN 在 MNIST、CIFAR10、CeUN 四个数据集上进行了基础学习能力、模型间 GAM 自比较、模型内 GAM 自比较、初始得分比较、图像生成可视化、时间效率比较等 6 组实验，采用生成对抗矩阵(Generative Adversarial Metric，GAM)和起始分数(Inception Scores)进行评估，结果表明，本文提出的 WIRGAN、GP-WIRGAN 具有良好的稳定性，可以生成高质量的图像.

关键词 图像生成；生成对抗网络；Wasserstein 距离；深度学习；权重剪枝；梯度惩罚

中图法分类号 TP18 **DOI 号** 10.11897/SP.J.1016.2020.00190

GP-WIRGAN: A Novel Image Recurrent Generative Adversarial Network Model Based on Wasserstein and Gradient Penalty

FENG Yong[1),2)] ZHANG Chun-Ping[1),2)] QIANG Bao-Hua[3),4)] ZHANG Yi-Yang[1),2)] SHANG Jia-Xing[1),2)]

[1)](*College of Computer Science, Chongqing University, Chongqing* 400044)

[2)](*Key Laboratory of Dependable Service Computing in Cyber Physical Society, Ministry of Education, Chongqing University, Chongqing* 400044)

[3)](*Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin, Guangxi* 541004)

[4)](*Guangxi Key Laboratory of Optoelectronic Information Processing, Guilin University of Electronic Technology, Guilin, Guangxi* 541004)

Abstract Most image generation models use a one-time image generation method, which obtains output through a single forward of generation model. But in practice, for example, painters usually repeatedly modify their paintings from coarse to fine during their creation time, which is a multi-stage

收稿日期：2018-08-21；在线出版日期：2019-06-12. 本课题得到国家自然科学基金(61762025)、国家重点研发计划(2017YFB1402400)、重庆市基础与前沿研究计划(cstc2017jcyjAX0340)、广西可信软件重点实验室开放课题(kx201701)、广西光电信息处理重点实验室(培育基地)基金(GD18202)、重庆市重点产业共性关键技术创新专项(cstc2017zdcy-zdyxx0047)、重庆市社会事业与民生保障科技创新专项(cstc2017shmsA20013)资助.**冯 永**，博士，教授，主要研究领域为大数据分析与数据挖掘、人工智能与大数据处理、深度学习与大数据检索. E-mail：fengyong@cqu.edu.cn. **张春平**，硕士，主要研究方向为深度学习与图像检索. **强保华**，博士，教授，主要研究领域为大数据处理与信息检索. **张逸扬**，硕士，主要研究方向为深度学习与大数据检索. **尚家兴**，博士，副教授，主要研究方向为人工智能与大数据处理.

process. Generative model reduces the manual marking requirements on image data, and can understand semantic meaning of the images well. The generative model can synthesize approximate real data from its learned data distribution. One of the main stream generative model is called Generative Adversarial Network (GAN). By utilizing game theory and deep learning, we can ultimately synthesize high-grade data samples based on two types of networks called generator and discriminator inside GAN model. GAN is well known for generating images, but has difficulty in training stably due to the irrational distance metric in optimizing target, which results in poorly generated sample diversity. Besides, most generative models generate images at a single cycle, but in fact, when the painter paints, he completes a painting on the basis of previous modifications. In order to guarantee the quality of the generated image and enhance the generation of sample diversity and the semantics of the sample, we simulate the process of repeating iterations and multiple modifications by the artist during painting, and generate samples using method we called "multi-generation". We chose Wasserstein distance to measure the distance between the real data distribution and the generated data distribution, proposed a framework named Wasserstein Image Recurrent Generative Adversarial Networks (WIRGAN). WIRGAN defines a generative model and a discriminative model, the generative model is used to gradually generate images, which consists of a recurrent feedback loop structure and can handle a time step parameter T of generation to control the complexity of model. Sample generated at time t is combined with the output of time $t-1$ by simply adding together, the generator takes the image generated from the last time step as output. The discriminator model is also constructed by a neural network, combining weight clipping to determine whether the input image is generated or true. WIRGAN uses Wasserstein distance as cost function, which aims to decrease the discrepancy between synthesized samples and real samples, training WIRGAN in an adversarial way. In addition, gradient penalty is also used in this paper to deal training difficulty that produced by weight clipping in WIRGAN. We further propose a Gradient Penalty Optimized Wasserstein Image Recurrent Generative Adversarial Networks Model (GP-WIRGAN). Finally, we adopt Generative Adversarial Metric (GAM) and inception score to evaluate the performance of our models on the quality and diversity of the generated samples. WIRGAN and GP-WIRGAN conducted five sets of comparative experiments on four datasets including MNIST, CIFAR10, CelebA and LSUN, which are the basic learning abilities comparison, the GAM comparisons within the model, the GAM comparisons between the models, the inception score comparisons, visualization, Time efficiency comparison. Extensive experiments show the proposed model has achieved good results in both evaluation criteria, which identify that WIRGAN and GP-WIRGAN has good stability and can generate high quantity images.

Keywords　image generating; generative adversarial networks; Wasserstein distance; deep learning; weight clipping; gradient penalty

1　引　言

图像可以表达出更直观、生动、高效的语义信息，逐渐成为日常生活中必不可少的表现方式，是海量大数据时代一种极为宝贵的媒体数据. 虽然互联网每天能产生超亿级的图像，但做到真正理解这些数据，从中挖掘出有价值的信息并加以利用的研究却很少，这样的环境造就了人工智能的快速发展. 人工智能领域经历了感知阶段和认知阶段两个时期. 在感知阶段中，感知机首先从外界接受各种信号，并依据这些信号做出判断，诸如语音识别、图像识别等研究领域，而在认知阶段中，机器能够对感知到东西形成自己的一套思维方式，能够根据一些演算做出

相应的决策，而不是单纯、机械地做出判断[1]. 无论是哪一阶段，理解数据本身才是最重要的.

理解是人类进行艺术创造的必经过程，艺术家作画时，通常需要不断叠加，反复修改后才能得到比较满意的作品，这些就是画家对自己作品逐渐深入理解的过程. 而理解对于人工智能来说，就是需要机器去挖掘已知数据中的普遍规律，并能将这些规律加以运用甚至再创造. 目前，生成模型是人工智能领域最典型的理解再创造的无监督模型，能够根据采样出的数据分布进行假设学习，通过学习得到样本分布的参数，训练得到拟合数据分布的生成模型，并应用生成模型生成全新的样本. 生成对抗网络(Generative Adversarial Networks，GAN)[2]作为典型的生成模型，能够学习数据里的潜在表示，并且已经在很多领域得到了应用，如图像处理、语音识别、异常检测、检索等，但 GAN 却难于训练，常常得到一些没有意义或低质量的采样结果，这就限制了中间结果可视化，妨碍了研究人员理解 GAN 学到的信息，且难以推进后续研究.

本文将循环结构和 Wasserstein 距离加入到经典的 GAN 网络中，由此提出了基于 Wasserstein 距离的循环生成对抗网络模型，简称为 WIRGAN. 和 GAN 相同，WIRGAN 模型也包括生成器和判别器两部分，其中，生成器采用循环结构，和 GRAN 中一样，但不叠加每一时刻的生成图像作为整个生成器的输出，而是直接将最后时刻的生成图像作为整个生成器的输出. 另外，WIRGAN 引入了 Lipschitz 约束，采用 Wasserstein 距离指导整个模型的训练，在判别器中使用权重剪枝技术，它做的是回归任务，而原始 GAN 中的判别器做的是二分类任务. 由于权重剪枝会引导模型学习简单的函数，从而影响生成图像的质量，所以采用梯度惩罚来代替权重剪枝，提出改进模型 GP-WIRGAN，输出结果的稳定性得到提高，并且能生成质量更高的图像.

2 相关工作

近年来，主流的生成模型包括生成对抗网络和变分自编码器(Variational Auto-Encoder，VAE)[3]两种.

基于变分自编码器的生成模型对数据进行假设，包括显式或隐式数据变量，然后训练得到一个自编码器，它由真实的训练数据训练，采用最大似然估计、近似法、马尔科夫链等方法，使它的潜在表示满足某种特定的概率分布，从而可以采样得到和数据分布一样的样本. 具体来说，Vincent 等人[4]通过在训练样本中加入随机噪声来重构图像；Rezende 等人[5]将隐变量概率模型和神经网络相结合，得到了一个非线性隐含高斯模型，并且同时优化变分参数和模型参数来解决可伸缩变分估计的问题；Gregor 等人[6]在 VAE 基础上，融合了注意力机制，提出 DRAW 网络，经过多次迭代生成近似于真实图像的生成图像数据. 但事实上，这些方法更偏向于从人的角度去理解数据，学习到的分布对于机器来说却有很大的限制，直接导致这些方法仅在简单数据集如(MNIST、NORB)上才能得到比较满意的样本，用在复杂数据集上，效果并不理想.

生成对抗网络是 Goodfellow 于 2014 年首次提出的，主要采用其对抗思想. GAN 包括生成模型和判别模型两部分，生成模型用来学习实际数据的潜在分布，并将学习到的信息用于生成新的数据样本，它也被称作生成器；判别模型用于分辨输入数据是真实数据还是生成数据，它也被称作判别器. 通过生成器和判别器进行共同学习训练，当学习达到纳什平衡[7]时，认为生成器已经估测到了真实数据的分布情况.

GAN 在图像生成领域影响深远. 在后来的研究中，许多研究人员以 GAN 为原型，在它的基础上作优化实验. Mirza 等人[8]限定了原始的输入，设置了条件约束，它可以是任何有利于生成任务的信息，通常为数据标签；Denton 等人[9]通过融合拉普拉斯金字塔模型和生成对抗网络，得到了从粗糙到精细的图像生成过程；Radford 等人[10]微调了生成对抗网络模型，提出了比原始的 GAN 更容易训练的 DCGAN，它的优势是不容易发生模式崩塌，即训练样本具有多样性，但生成图像样本单一的现象；Chen 等人[11]结合信息论和生成对抗网络，提出了 infoGAN 模型；Im 等人[12]使用循环结构替换了生成器，提出了 GRAN 模型；Zhao 等人[13]将能量函数引入到生成对抗网络的学习当中，让生成图像拥有更高的能量值，真实图像拥有更低的能量值，判别器通过能量函数来判断输入是否为生成数据；在生成对抗网络提出不久后，人们发现训练困难是 GAN 模型一个很大的问题，Arjovsky 等人针对该问题，总结出了相关的数学证明，采用 Wasserstein 距离来度量真实分布和生成分布之间的距离，引入权重减枝来约束判别模型，提出了 WGAN[14]；由于权重减枝极易诱导模型学习到两极化参数，影响生成质

量，Gulrajani 等人用梯度惩罚代替权重减枝，提出了改进模型 WGAN-GP[15]，提高了模型的性能. 在近期研究中，Karras 等人[16]将生成模型和判别模型设计为逐层增长的形式，在多种数据集上均得到不错的结果. 此外，GAN 也被运用到其他领域，Peng 等人[17]将其运用到跨媒体检索；百度将 GAN 运用到语音识别[18].

然而，大多数基于神经网络的生成对抗模型都是采用一次性的方式来生成样本，这也使得生成图像的像素点以单一潜在分布为条件[15-16,18]. 因此我们结合深度学习，通过不断地训练学习，解决生成图像存在的信息缺失等问题，从而实现深层次理解语义信息，使得生成模型能够生成近似于真实图像的样本.

3　模型框架

本模型对图像生成进行研究，结合 WGAN 和 WGAN-GP 模型的处理思路，改进 GRAN 模型. 我们从两方面入手：一是生成模型由多个结构相同的神经网络结构组成，前一时刻的输出作为后一时刻的输入，将最终时刻的输出作为生成模型的整体输出；二是选用 Wasserstein 距离作为生成分布和真实分布之间的距离衡量方法. 本文提出了 Wasserstein 图像循环生成对抗网络(WIRGAN). 并且，对模型进行优化，提出改进模型 GP-WIRGAN.

3.1　WIRGAN

3.1.1　GRAN 的一般性

GRAN 的损失函数和原始 GAN 一样，都是极大极小对抗方式，具体定义如式(1)：

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

这里 D 表示判别器，G 表示生成器. 由于 G 和 D 是交替训练的，所以训练 D 时，需要固定 G. 假设 G 和 D 都有强大的学习能力，求得当前生成器对应的最优判别器 $D_G^*(x)$，如式(2)所示：

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

在最优判别器 $D_G^*(x)$ 的情况下训练 G，此时 G 的损失函数可以等价变换成式(3)的形式：

$$\begin{aligned}C(G)&=\max_D V(G,D)\\&=\mathbb{E}_{x\sim p_{\text{data}}}[\log D_G^*(x)]+\mathbb{E}_{z\sim p_z}[\log(1-D_G^*(G(x)))]\\&=\mathbb{E}_{x\sim p_{\text{data}}}[\log D_G^*(x)]+\mathbb{E}_{x\sim p_g}[\log(1-D_G^*(x))]\\&=-\log(4)+2\cdot D_{\text{JS}}(p_{\text{data}}\|p_g)\end{aligned} \tag{3}$$

此时生成器的任务相当于最小化真实分布 p_{data} 和生成分布 p_g 之间的 JS 散度. 当且仅当 $p_g=p_{\text{data}}$，$C(G)$取得最小值. 但是基于两个前提：其一是 G 和 D 都有足够强大的学习优化能力，能够对自身进行不断优化，各自提高自己的生成能力和判别能力，从而达到各自最优状态；其二是数据的生成分布和真实分布之间存在重叠区域. 但实际上，生成器 G 与判别器 D 的学习能力存在差异，一般情况下 D 的学习能力要比 G 更为强大，这就会导致 G 难以从 D 的训练结果中获取有用的反馈信息来对自身的生成样本能力进行优化，从而使得生成分布 p_g 和真实分布 p_{data} 之间缺少重叠区域，或者重叠可以忽略，这时生成器的损失值为常数 $-\log 2$，梯度消失，G 无法对参数进行更新，导致模型不稳定，生成的样本缺乏多样性.

3.1.2　目标函数

正如 WGAN 模型中的分析，JS 散度在高维空间并不能为模型提供有意义的梯度，但 Wasserstein 距离却能做到这一点. 所以，本文提出了 Wasserstein 图像循环生成对抗网络模型——WIRGAN. 将 Wasserstein 距离作为生成分布 p_g 和真实分布 p_{data} 之间距离衡量的标准. 除此之外，因为引入 K-Lipschitz 连续来限定距离函数中参数值变动范围，所以要求训练模型时网络参数变动幅度要约束在一个范围以内，即参数在每一次更新时变动幅度不能超过某常量. 因此真实数据分布 p_{data} 和生成数据分布 p_g 之间的 Wasserstein 距离可以表示为

$$D_w=\mathbb{E}_{x\sim p_{\text{data}}}[f_w(x)]-\mathbb{E}_{x\sim p_g}[f_w(x)] \tag{4}$$

上式中，D_w 越小则生成分布 p_g 就越可能接近真实分布 p_g. 另外，由于 Wasserstein 距离引入了 K-Lipschitz 连续对其进行约束，使其在任何时候的函数都是可微的，从而解决了 GAN 模型训练过程中梯度消失问题，所以得到 WIRGAN 判别器的目标函数：

$$obj^D=\min(\mathbb{E}_{x\sim p_g}[f_w(x)]-\mathbb{E}_{x\sim p_{\text{data}}}[f_w(x)]) \tag{5}$$

由于生成模型和判别模型是“你失我得”的关系，所以有：

$$obj^G=-obj^D \tag{6}$$

又因为式(5)中的$\mathbb{E}_{x\sim p_{\text{data}}}[f_w(x)]$与生成器无关，所以可以得到生成器的目标函数为

$$obj^G=\min(-\mathbb{E}_{x\sim p_g}[f_w(x)]) \tag{7}$$

3.1.3　模型结构

Wasserstein 图像循环生成对抗网络(WIRGAN)包含生成模型与判别模型两部分. 图 1 是 WIRGAN

的整体框架.其中,生成模型由多个结构相同的神经网络组成,模拟了人类理解场景的过程,这和循环生成对抗网络模型(GRAN)中的生成模型相似,但是也存在不同之处,主要有四点:(1)WIRGAN 将最后一个时间步骤的输出作为最终生成样本,而 GRAN 是将所有时间步骤的输出进行叠加计算后作为最终输出样本;(2)GRAN 比 WIRGAN 多一个全联接层,通过 tanh 函数将随机向量变换到[−1,1]之间;(3)GRAN 的判别器实现的是一个二分类任务,它只判断输入数据是否来自真实样本,因此需要在模型中使用 sigmoid 函数;但是 WIRGAN 损失函数的构建是为了对真实数据分布与生成数据分布进行回归分析,尽可能的拟合它们之间的相似度,而不是如 GRAN 进行分类;(4)WIRGAN 对判别模型使用了权重剪枝技术.下面将对生成模型和判别模型分别进行介绍.

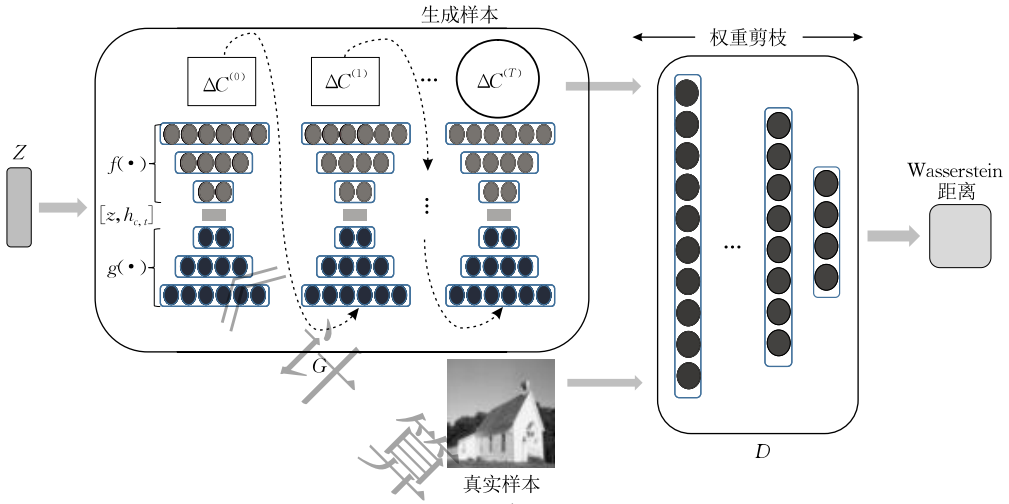

图 1 WIRGAN 框架

(1)生成模型

WIRGAN 的生成模型的设计思路与画家作画过程类似:画家通过对画作不断的进行迭代修改,反复添加内容,从而得到最终成品.该生成器由多个结构相同的模型组成,每一个单独结构为一个时刻,后一时刻的图像都是在前一时刻的基础上生成的,将最后时刻的生成图像作为整个生成模型的输出.

图 2 显示了单个结构,包含 $g(\cdot)$和 $f(\cdot)$两部分.与真正的自编码器相比,不同之处在于:(1)当 $t=1$ 时,由于没有来自前一时间步骤的生成图像作为输入,所以此时刻的结构没有编码器,只有解码器,所以模型从 $f(\cdot)$开始计算,这和 GRAN 中的结构相同,但有别于传统自编码器;(2)$T>1$ 时,t 时刻,解码器 $f(\cdot)$会将编码器 $g(\cdot)$的结果和噪声向量进行联合,再进行解码计算,而非传统自编码器直接将编码器的结果进行解码计算,所以模型在 t 时刻的生成图像要比 $t-1$ 时刻的图像信息更丰富.以上两点说明了模型中的单个结构并不是还原图像,而是在前一个图像上增加细节.

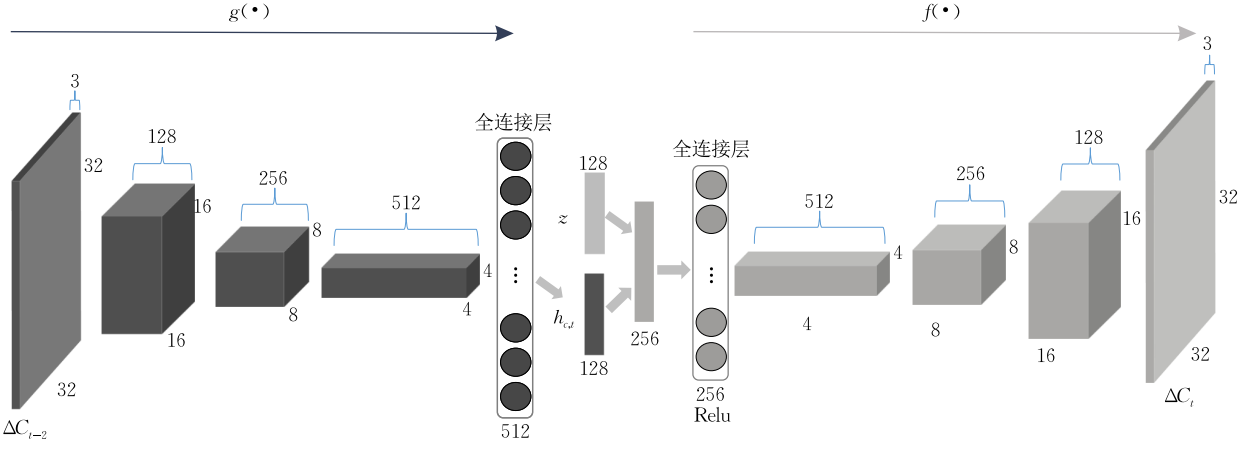

图 2 WIRGAN 的生成器的单个结构

$g(\cdot)$的前几层为卷积层，每个卷积层后面都跟随着一个归一层，对数据归一化后再进行线性变换改善数据分布，在卷积层中使用 ReLU 作为激活函数，最后一层为全连接层；$f(\cdot)$的第一层为全连接层，随后全是反卷积层，与 $g(\cdot)$类似，每一个反卷积层后面都连接着一个归一层. 除了最后一个反卷积层使用 tanh 作为激活函数外，其余反卷积层的激活函数都选择 ReLU. 结构上，$g(\cdot)$和 $f(\cdot)$是对称的.

不同于 GRAN，WIRGAN 的生成器输入 z_t 为从标准正态分布中采样出的 128 维、范围为$[-1,1]$的向量，并不需要通过一个使用 tanh 作为激活函数的全连接层将输入数据变换为$[-1,1]$，减少了计算量. 对于每一个时间步骤 $t=1,2,\cdots,T$，$\mathcal{C}$ 为生成器的输出，整个计算过程如下：

$$z_t \sim p(Z) \tag{8}$$

$$h_{c,t}=g(\Delta C_{t-1}) \tag{9}$$

$$\Delta C_t=f([z_t,h_{c,t}]) \tag{10}$$

循环结构的计算从编码器 $g(\cdot)$开始，$g(\cdot)$对前一个时间步骤 $t-1$ 的生成样本 ΔC_{t-1} 进行编码，得到 $h_{c,t}$，然后，联合 $h_{c,t}$ 和从标准正态分布取样的 128 维的噪声 z_t 得到$[z_t,h_{c,t}]$，以得到将其输入到解码器 $f(\cdot)$中生成当前时间步骤 t 的样本 ΔC_t. 生成模型的结构大小由次数 T 控制，T 可为任何整数，本模型设置 $T=\{1,3,5\}$. 当 $t=1$ 时，没有来自前一时间步骤的生成图像作为输入，所以此时刻的结构中没有编码器，只需将 $\boldsymbol{h}_{c,0}$ 初始化为零向量，这和 GRAN 中的结构相同.

不同于 GRAN，WIRGAN 的生成器的最终输出为最后一次时间步骤生成的图像，即：

$$\mathcal{C}=\Delta C_T \tag{11}$$

(2) 判别器

判别器的结构如图 3 所示，判别器主要由卷积层与全连接层构成，相对于生成器而言其结构比较简单. 选择 LeakyReLU 作为每一个卷积层的激活函数，最后一个全连接层输出 Wasserstein 距离. 使用权重剪枝，让参数每次更新幅度都在常数 $-c$ 与 c 之间.

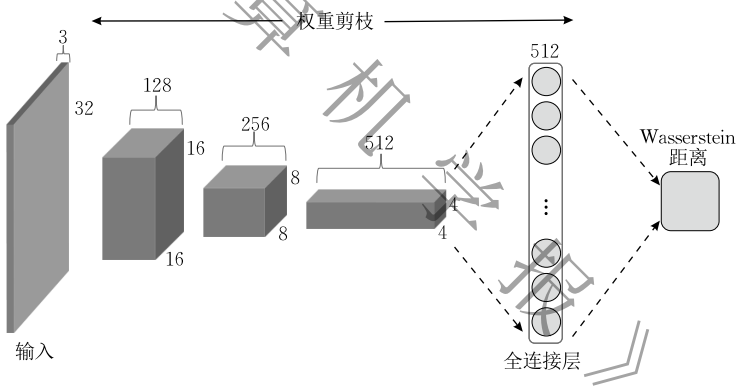

图 3　WIRGAN 的判别器

3.1.4　训练流程

WIRGAN 中的生成模型是一个循环结构，初始计算 $t=1$ 时，$g(\cdot)$没有前一刻生成图像作为输入，所以设置 $g(\cdot)$在时刻 $t=0$ 的输出 $\boldsymbol{h}_{c,0}$ 为 128 维的零向量，如算法 1 所示.

算法 1.　WIRGAN 生成样本的算法.

输入：随机噪声 z，隐藏层初始参数 $\boldsymbol{h}_{c,0}$，时间步骤 $T=\{1,3,5\}$

输出：生成样本 $\mathcal{C}$

1. WHILE $t<T$ DO
2. 　　$z_t \sim p(z)$
3. 　　$h_{c,t}=g(\Delta C_{t-1})$
4. 　　$\Delta C_{t-1}=f([z_t,h_{c,t}])$
5. END WHILE
6. $\mathcal{C}=\Delta C_T$

在实际的对抗训练中，WIRGAN 选用 RMSProp 算法对模型中的参数进行优化. WIRGAN 进行批量训练，判别模型更新多次，生成模型才更新一次.

使用 w 表示判别器中的参数值，常数 c 表示在判别器中对参数更新的约束范围，θ 表示生成器中参数. 则可用下述算法 2 描述 WIRGAN 具体训练过程.

算法 2.　WIRGAN 算法.

输入：随机噪声 z，真实样本 x，生成器参数 θ，判别器参数 w，权重剪枝 c，学习率 ℓ，批处理 m，判别器更新次数 n_D

输出：Wasserstein 距离

1. WHILE θ has not converged DO
2. 　FOR $t=0,\cdots,n_D$ DO
3. 　　Samples of real data $\{x^{(i)}\}_{i=1}^{m} \sim p_{\text{data}}(x)$
4. 　　Samples of latent variables $\{z^{(i)}\}_{i=1}^{m} \sim p_g(z)$
5. 　　$D_w \leftarrow$

$$\nabla_w\left[-\frac{1}{m}\sum_{i=1}^{m}D_w(x^{(i)})+\frac{1}{m}\sum_{i=1}^{m}D_w(G_\theta(z^{(i)}))\right]$$

6. $w\leftarrow w+\ell\cdot\mathrm{RMPSProp}(w,D_w)$
7. $w\leftarrow clip(w,-c,c)$
8. END FOR
9. Samples of latent variables $\{z^{(i)}\}_{i=1}^{m}\sim p_g(z)$
10. $G_\theta\leftarrow-\nabla_\theta\frac{1}{m}\sum_{i=1}^{m}D_w(G_\theta(z^{(i)}))$
11. $\theta\leftarrow\theta-\ell\cdot\mathrm{RMPSProp}(\theta,G_\theta)$
12. END WHILE

3.2 GP-WIRGAN

权重减枝导致 WIRGAN 模型优化困难，主要体现在两个方面：

(1) 权重剪枝将判别器的参数限制在一个范围内(如$[-0.01,0.01]$)，使判别器在训练过程中极易学习到两极化的参数，这意味着在数据分布上，判别器只能捕捉到低阶矩，使得神经网络拟合能力被浪费，并且直接影响到生成图像的质量.

(2) 在 WIRGAN 中，生成器和判别器的神经网络模型都是由多层构成，因此在对权重的剪枝范围进行设置时，不合适的范围设置在经过若干层网络传输放大后都有可能造成指数衰减或爆炸.

3.2.1 目标函数

正如 WGAN-GP 中计算分析的一样，由于权重剪枝引导模型判别学习简单的函数来满足 K-Lipschitz 约束，不足以捕捉到数据的高阶矩，影响生成图像的质量.

我们采用梯度惩罚进一步优化模型，提出改进模型 GP-WIRGAN. 我们使用随机采样的方法获取真样本 x_{data}、假样本 x_g 以及一个随机数 ϵ，范围为$[0,1]$：

$$x_{\text{data}}\sim p_{\text{data}},x_g\sim p_g,\epsilon\sim uniform[0,1] \quad (12)$$

然后在 x_{data} 和 x_g 的连线上随机插值采样：

$$\hat{x}=\epsilon x_{\text{data}}+(1-\epsilon)x_g \quad (13)$$

将 $\hat{x}$ 所满足的分布记为 $p_{\hat{x}}$，最终得到改进模型梯度惩罚优化的 Wasserstein 图像循环生成对抗网络模型(GP-WIRGAN)的目标函数为

$$obj^{(G,D)}=\mathbb{E}_{\tilde{x}\sim p_g}[D_w(\tilde{x})]-\mathbb{E}_{x\sim p_{\text{data}}}[D_w(x)]+\lambda\mathbb{E}_{\hat{x}\sim p_{\hat{x}}}[\|\nabla_{\hat{x}}D_w(\hat{x})\|_2-1]^2 \quad (14)$$

这时，便可得到判别器的目标函数：

$$obj^{D}=\min(\mathbb{E}_{\tilde{x}\sim p_g}[D_w(\tilde{x})]-\mathbb{E}_{x\sim p_{\text{data}}}[D_w(x)]+\lambda\mathbb{E}_{\hat{x}\sim p_{\hat{x}}}[\|\nabla_{\hat{x}}D_w(\hat{x})\|_2-1]^2) \quad (15)$$

生成器的目标函数为

$$obj^{G}=\min(-\mathbb{E}_{\tilde{x}\sim p_g}[D_w(\tilde{x})]) \quad (16)$$

有一点值得说明，GAN 中纳什平衡是建立在模型做二分类任务的基础上分析的，并且是假设生成模型和判别模型都有足够的能力时达到的一种理想情况，而 WIRGAN 和 GP-WIRGAN 由于是用 Wasserstein 距离作为目标函数，属于回归任务，同时模型的能力达不到理想状态，所以无法判断是否是接近纳什平衡.

3.2.2 训练流程

GP-WIRGAN 通过独立地对每个样本应用约束来实现梯度惩罚的效果. 如果在 GP-WIRGAN 中的判别模型中使用批正则化层会使得输入数据之间建立依赖，在梯度下降时对样本梯度改变造成影响. 所以，GP-WIRGAN 的结构在 WIRGAN 模型的基础上进行了微调，生成模型的结构保持不变，判别模型中取消 Batch Normalization 层，并且去掉了权重剪枝技术. 为了在训练时取得相对较好的生成样本，WIRGAN 采用了 RMSProp 算法，但是 GP-WIRGAN 中优化算法是不需要小心采用的，实验在对参数进行更新时采用 Adam 优化算法，由于这是非确定性算法，还是得根据实验效果进行验证. 具体的训练流程如算法 3 所示.

算法 3. GP-WIRGAN 的算法.

输入：随机噪声 z，真实样本 x，生成器参数 θ，判别器参数 w，梯度惩罚系数 λ，学习率 ℓ，批处理 m，判别器更新次数 n_D，Adam 超参 β_1,β_2

输出：Wasserstein 距离

1. WHILE θ has not converged DO
2. FOR $t=0,\cdots,n_D$ DO
3. FOR $i=1,\cdots,m$ DO
4. Sample from the real data $x\sim p_{\text{data}}$
5. Sample from the latent variable $z\sim p_z$
6. A random number $\epsilon\sim uniform[0,1]$
7. $\tilde{x}\leftarrow G_\theta(z)$
8. $\hat{x}\leftarrow\epsilon x+(1-\epsilon)\tilde{x}$
9. $obj_{(i)}^{D}\leftarrow$ $[D_w(\tilde{x})-D_w(x)+\lambda(\|\nabla_{\hat{x}}D_w(\hat{x})\|_2-1)^2]$
10. END FOR
11. $w\leftarrow\mathrm{Adam}\left(\nabla_w\frac{1}{m}\sum_{i=1}^{m}obj_{(i)}^{D},w,\ell,\beta_1,\beta_2\right)$
12. END FOR
13. Samples from the latent variables $z\sim p_z$
14. $\theta\leftarrow\mathrm{Adam}\left(\nabla_\theta\frac{1}{m}\sum_{i=1}^{m}-D_w(G_\theta(z)),\theta,\ell,\beta_1,\beta_2\right)$
15. END WHILE

4 实验及结果分析

4.1 数据集

本实验用了 4 个公开数据集：MNIST、CIFAR10、

LSUN、CelebA. MNIST 是手写体数据集[19]，一共有 7 万张图片，0～9 中的所有数字都包含在其中，其中训练集包含 6 万张图像，测试集包含 1 万张图像，每张图片大小为 28×28；CIFAR10 数据集[20]共有 10 类，包含了 6 万张 32×32 大小的图像，其中训练集包含 5 万张图像，测试集包含 1 万张图像；LSUN[21]是自然场景数据集，包含 10 种不同的场景，在本实验中，用户外教堂这一单个数据集进行训练，其中，包含 126 227 张训练集图像，1000 张测试集图像，将每张图像处理为 64×64 大小；CelebA (CelebFaces Attributes Dataset)[22]是大型人脸图像数据集，包含 10 177 位名人的 202 599 张人脸图像，将图像处理为 128×128 大小，训练集包含 18 万张图像，测试集包含 22 599 张图像.

4.2　评价指标体系

为了说明 WIRGAN 和 GP-WIRGAN 模型的性能，实验中并没有采用重构误差对生成器里的编码器的编码结果进行单独评价，而是用了两个直接衡量指标. 为了评价生成样本质量，我们采用生成对抗测评(Generative Adversarial Metric，GAM)[12]这一指标；为了评价生成样本多样性，我们采用初始得分(Inception score)[23]这一指标. 下面将对这两个指标进行具体介绍.

(1) GAM

GAM 包含 GAN 的训练阶段和测试阶段，假设 $M_1=\{(G_1,D_1)\}$ 和 $M_2=\{(G_2,D_2)\}$ 为两个不同的生成对抗网络模型，训练阶段分别对 M_1 和 M_2 进行训练，测试阶段分别用对方的判别器来判别自身生成器生成的样本.

用同一个随机噪声 z 作为 M_1 和 M_2 中生成器的输入，分别用 $G_1(z)$ 和 $G_2(z)$ 表示 M_1 和 M_2 生成的样本，用 x_{train} 和 x_{test} 表示训练阶段和测试阶段输入到判别器中的真实数据，两个 GAN 之间的 GAM 比较数据记录如表 1 所示.

表 1　GANs 之间的 GAM 比较

	M_1	M_2
M_1	$D_1(G_1(z)),D_1(x_{\text{train}})$	$D_1(G_2(z)),D_1(x_{\text{test}})$
M_2	$D_2(G_1(z)),D_2(x_{\text{test}})$	$D_2(G_2(z)),D_2(x_{\text{train}})$

在GAM比较中，有两个重要的比值：r_{test} 和 r_{sample}，$err(\cdot)$ 表示分类错误率，具体定义如下：

$$r_{\text{test}}=\frac{err(D_1(x_{\text{test}}))}{err(D_2(x_{\text{test}}))} \tag{17}$$

$$r_{\text{sample}}=\frac{err(D_1(G_2(z)))}{err(D_2(G_1(z)))} \tag{18}$$

式(17)中 r_{test} 能够体现 M_1 和 M_2 中的泛化能力，它是基于测试数据进行的判别. 判别器过拟合将会使生成器生成的样本与真实图像差异较大. r_{sample} 表示哪一个模型生成的样本更容易“欺骗”另一个模型中的判别器，即生成的样本质量更高. 为了减少判别器在不同模型中的对数据的偏向性，用 r_{test} 来保证比较的公平性，用 r_{sample} 来决定胜出模型. 那么具体的判别规则如下：

$$winner=\begin{cases}M_1, & r_{\text{sample}}<1 \text{ and } r_{\text{test}}\approx 1\\ M_2, & r_{\text{sample}}>1 \text{ and } r_{\text{test}}\approx 1\\ Tie, & \text{其他}\end{cases} \tag{19}$$

为了避免两个模型判别器过拟合，则需要 $r_{\text{test}}\approx 1$，若 r_{test} 离 1 非常远，这时进行的 GAM 比较是不公平的，因为此时其中一个判别器对数据存在过拟合现象. 在本文实验中，若 $0.85<r_{\text{test}}<1.1$，则认为 $r_{\text{test}}\approx 1$.

(2) 初始得分

为了获得生成样本的全局和局部特征，我们采用初始得分. 初始得分越高，代表生成样本的多样性越高；反之，多样性越低.

4.3　实　验

4.3.1　实验结构设置

在实验中，针对 4 个不同的数据集，我们设置的 WIRGAN 和 GP-WIRGAN 对应的结构不同，表 2 展示了具体模型卷积核的结构设置. 选定卷积后的特征图大小不小于 2×2，以确保特征图中包含的信息充分，同时选定神经元基数(即特征图基数)为 128. 生成器和判别器中的卷积层层数以及每一层的个数都是根据输入图像的大小决定的. 以 CIFAR10 为例，它是彩色图象，其输入大小是 32×32×3，经过卷积之后特征图的大小变化为 32×32→16×16→8×8→4×4，对应特征图数量为 3→128→256→512，则对应到 $g(\cdot)$ 和判别器 D 中就有三层卷积层，$f(\cdot)$ 中就有三层反卷积层. 表 2 展示了两个模型中生成器和判别器在不同数据集中特征图的通道数. 同时，我们采用标准正态初始化方法对两个模型的权值进行初始化.

表 2 不同数据集对应的模型卷积核结构

数据集	大小	$f(\cdot)$中的通道数	$g(\cdot)$中的通道数	D 中的通道数
MNIST	28×28	128→256	256→128	128→256
CIFAR10	32×32	128→256→512	512→256→128	128→256→512
LSUN	64×64	128→256→512→1024	1024→512→256→128	128→256→512→1024
CelebA	64×64	128→256→512→1024	1024→512→256→128	128→256→512→1024

4.3.2 simWIRGAN vs. simGP-WIRGAN

本实验是为了解释使用权重剪枝的 WIRGAN 和使用梯度惩罚的 GP-WIRGAN 学习数据的情况.

我们构建了简单的 WIRGAN 和 GP-WIRGAN 模型,分别表示为simWIRGAN和simGP-WIRGAN,两者结构相同,目标函数不同.生成器中包含结构对称的两部分 $g(\cdot)$和 $f(\cdot)$,$f(\cdot)$由多个全连接层组成,每个全连接层均包含 512 个神经元并采用 Relu 激活函数,最后一层不使用激活函数,循环次数设置为 $T=\{1,3\}$.判别器的结构与生成器类似,也由多个使用 Relu 激活函数的全连接层组成,最后一层不使用激活函数.

在本实验中,两种模型都采用批量训练且不使用归一层,128 个采样数据点作为一个批次.用从 Swiss roll 与高斯噪声中采样得到的数据点代替真实数据,采用 3 种输入方案:首先是从 Swiss roll 点集中随机采样 128 个数据点;其次是从 8 个高斯噪声点集中随机采样 128 个数据点;最后是从 25 个高斯噪声点集中随机采样 128 个数据点.我们基于两种不同的生成样本,用 simWIRGAN 和 simGP-WIRGAN 进行了两组实验.

(1) 添加随机噪声的真实分布

该实验直接将加了随机噪声的真实数据作为生成样本.我们在 simWIRGAN 和 simGP-WIRGAN 中进行训练,根据不同的输入得到的结果如图 4 所示.浅灰色点代表真实数据,曲线代表模型学习到的数据分布的情况.两行图像分别为 simWIRGAN 和 simGP-WIRGAN 的训练结果.

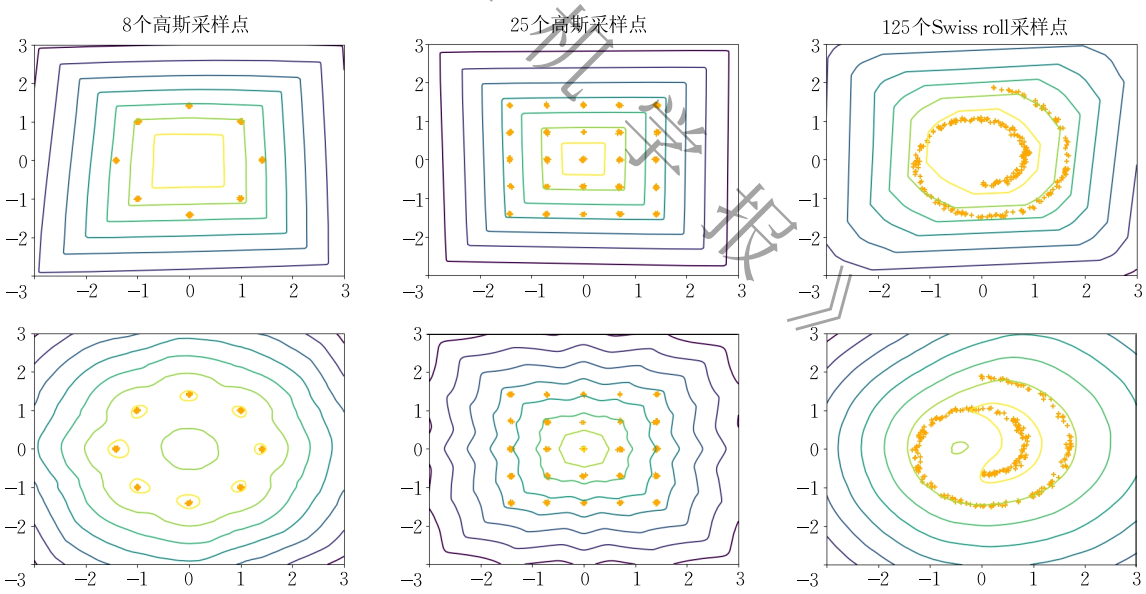

图 4 不经过生成器的模型学习结果(第一排为 simWIRGAN 训练结果,第二排为 simGP-WIRGAN 训练结果)

(2) simWIRGAN 和 simGP-WIRGAN 中生成器生成的数据

图 5 展示了在 $T=1$ 时刻三种不同输入的训练情况,图 6 展示了在 $T=3$ 时刻三种不同输入的训练情况,图中浅色点为真实数据,深色点为生成数据,不规则曲线为数据分布,两行分别为 simWIRGAN 与 simGP-WIRGAN 的训练结果.

对比两组实验可以看出,GP-WIRGAN 模型相对于 WIRGAN 模型学习到的曲线更多了一些弧度,在一定程度上解决了 WIRGAN 的优化困难问题.反映出了梯度惩罚方法使得 GP-WIRGAN 获取了更高阶的矩,拥有更好的数据特征.

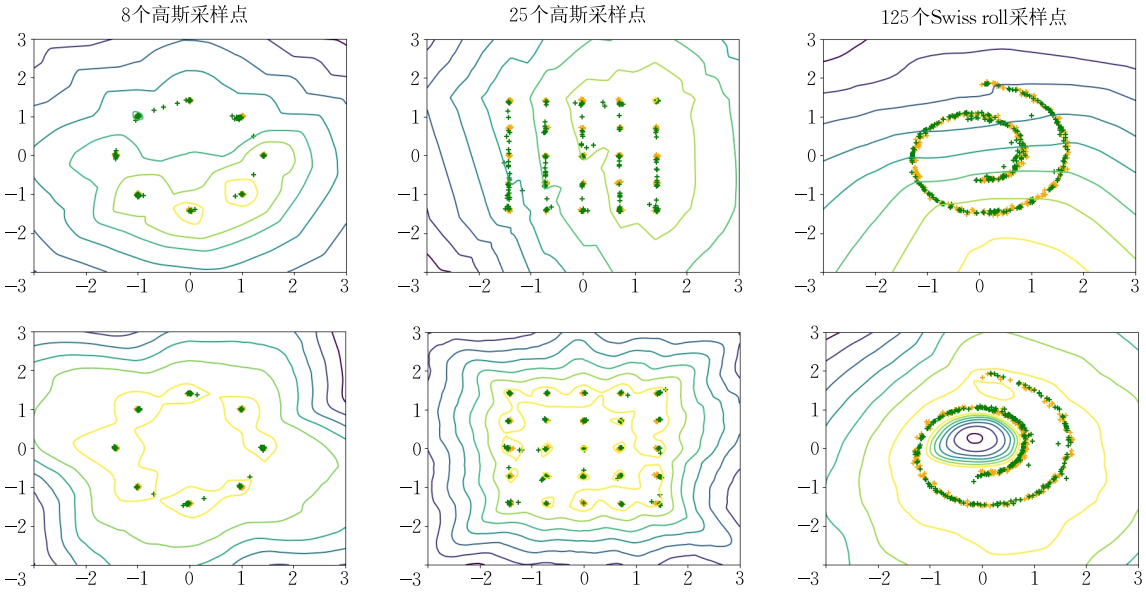

图 5　$T=1$ 时 simWIRGAN 和 simGP-WIRGAN 的学习结果(第一排为 simWIRGAN 训练结果,第二排为 simGP-WIRGAN 训练结果)

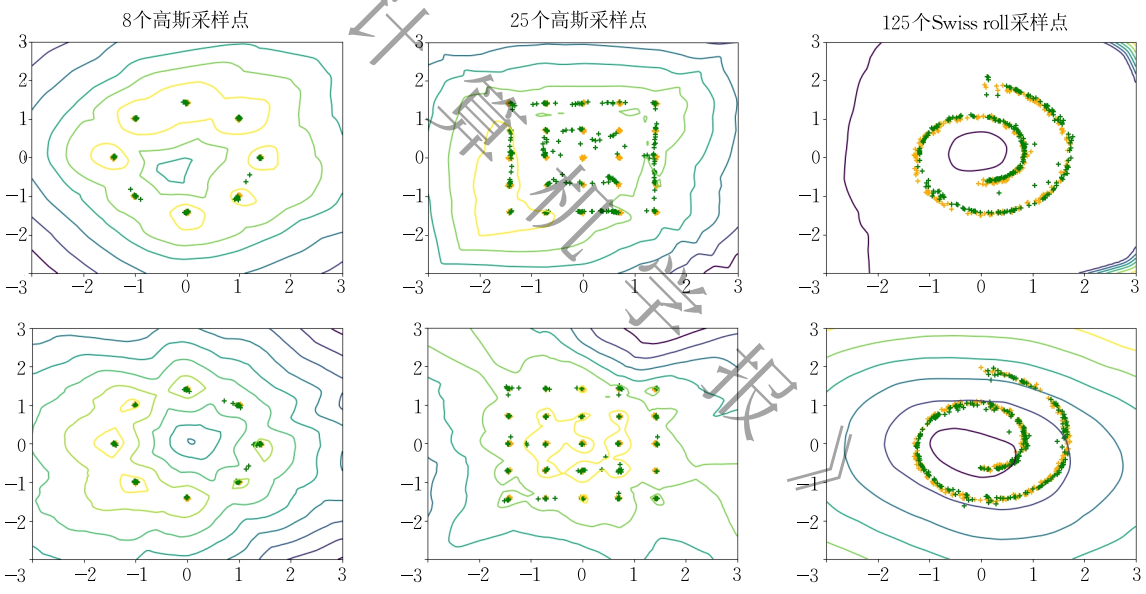

图 6　$T=3$ 时 simWIRGAN 和 simGP-WIRGAN 的学习结果(第一排为 simWIRGAN 训练结果,第二排为 simGP-WIRGAN 训练结果)

4.3.3　模型内的 GAM 自比较

因为 WIRGAN 模型的生成器由多个神经网络构成且每个神经网络的结构相同,所以会导致不同生成器生成的结果也会有所不同. 为了比较不同时间步骤下模型生成图像的质量,在模型内进行 GAM 自比较.

设置生成器的时间步骤 $T=\{1,3,5\}$,随机选取 1000 张测试集图片作为真实样本,生成样本为用自比较模型的生成器生成的 1000 张图片,然后进行模型间 GAM 比较. 在 GAM 对比中,前一个模型看作 $M_1=\{(G_1,D_1)\}$,后一个模型看作 $M_2=\{(G_2,D_2)\}$. 表 3 为 WIRGAN 的 GAM 自比较结果,表 4 为 GP-WIRGAN 的 GAM 自比较结果.

表 3　WIRGAN 的 GAM 自比较结果

Dataset	WIRGAN	GAM	r_{test}	r_{sample}
MNIST	$T=1$	$T=3$	0.95	1.003
	$T=1$	$T=5$	0.95	1.007
	$T=3$	$T=5$	0.99	1.581
CIFAR10	$T=1$	$T=3$	1.00	1.009
	$T=1$	$T=5$	1.00	1.001
	$T=3$	$T=5$	1.00	1.875
LSUN	$T=1$	$T=3$	1.08	1.331
	$T=1$	$T=5$	1.00	2.151
	$T=3$	$T=5$	1.03	17.81
CelebA	$T=1$	$T=3$	0.97	3.366
	$T=1$	$T=5$	0.99	3.214
	$T=3$	$T=5$	0.99	14.22

表 4 GP-WIRGAN 的 GAM 自比较结果

Dataset	GP-WIRGAN	GAM	r_{test}	r_{sample}
MNIST	$T=1$	$T=3$	1.00	1.245
	$T=1$	$T=5$	1.00	1.653
	$T=3$	$T=5$	1.00	1.008
CIFAR10	$T=1$	$T=3$	1.00	1.018
	$T=1$	$T=5$	1.00	1.754
	$T=3$	$T=5$	1.00	6.437
LSUN	$T=1$	$T=3$	1.00	5.822
	$T=1$	$T=5$	1.00	5.100
	$T=3$	$T=5$	1.00	1.821
CelebA	$T=1$	$T=3$	1.00	2.770
	$T=1$	$T=5$	1.00	3.362
	$T=3$	$T=5$	1.00	5.891

从表 3 和表 4 中可以看到，WIRGAN 和 GP-WIRGAN 的生成图像的质量随着时间步骤的越高而越好. 值得注意的是，当 $T=1$ 时，WIRGAN1 等价于 WGAN，GP-WIRGAN1 等价于 WGAN-GP，而当 $T>1$ 时，生成器则为一个循环结构，这也说明了利用循环结构进行“多次生成”，有利于提高生成图像的质量. 另外，在 WIRGAN 的自比较实验中，从实验数据可以看到，以 $T=1$ 时的 WIRGAN（等同于 WGAN）为参照基准，虽然 $T=3$ 和 $T=5$ 时的 WIRGAN 都是 winner，但是 $T=5$ 时相对于 $T=1$ 时的 WIRGAN（WGAN）的错误率只比 $T=3$ 时相对于 $T=1$ 时的 WIRGAN（WGAN）的错误率的低了一点，所以得出结论：$T=3$ 时 WIRGAN 的整体性能最高.

4.3.4 模型间的比较

模型间的比较包含两类比较：一组是循环结构的生成对抗模型（GRAN，WIRGAN，GP-WIRGAN）之间的两两 GAM 比较；另一组是和非生成对抗网络的生成模型的比较.

（1）和循环结构的生成对抗网络的比较

我们在 GRAN、WIRGAN、GP-WIRGAN 之间进行两两比较，因为我们的模型为循环结构. 表 5 为 WIRGAN 和 GP-WIRGAN 之间的不同时间步骤下的 GAM 比较结果；表 6 为 GRAN 和 WIRGAN 之间的不同时间步骤下的 GAM 比较结果；表 7 为 GRAN 和 GP-WIRGAN 之间的不同时间步骤下的 GAM 比较结果.

表 5 WIRGAN 和 GP-WIRGAN 之间的 GAM 比较结果

Dataset	WIRGAN	GP_WIRGAN	r_{test}	r_{sample}
MNIST	$T=1$	$T=1$	0.95	1.001
	$T=3$	$T=3$	0.95	1.006
	$T=5$	$T=5$	1.00	2.287
CIFAR10	$T=1$	$T=1$	1.00	1.638
	$T=3$	$T=3$	1.00	3.681
	$T=5$	$T=5$	1.00	2.568
LSUN	$T=1$	$T=1$	0.93	5.020
	$T=3$	$T=3$	0.98	1.498
	$T=5$	$T=5$	0.98	5.267
CelebA	$T=1$	$T=1$	0.96	7.152
	$T=3$	$T=3$	0.99	4.212
	$T=5$	$T=5$	0.99	3.354

表 6 GRAN 和 WIRGAN 之间的 GAM 比较结果

Dataset	GRAN	WIRGAN	r_{test}	r_{sample}
MNIST	$T=1$	$T=1$	0.94	1.684
	$T=3$	$T=3$	0.89	1.367
	$T=5$	$T=5$	0.92	6.585
CIFAR10	$T=1$	$T=1$	0.88	6.821
	$T=3$	$T=3$	0.91	1.525
	$T=5$	$T=5$	0.91	5.759
LSUN	$T=1$	$T=1$	0.89	1.398
	$T=3$	$T=3$	0.95	1.213
	$T=5$	$T=5$	0.90	3.993
CelebA	$T=1$	$T=1$	0.99	15.72
	$T=3$	$T=3$	0.99	10.21
	$T=5$	$T=5$	0.99	13.02

表 7 GRAN 和 GP-WIRGAN 之间的 GAM 比较结果

Dataset	GRAN	GP-WIRGAN	r_{test}	r_{sample}
MNIST	$T=1$	$T=1$	0.92	3.751
	$T=3$	$T=3$	0.89	2.683
	$T=5$	$T=5$	0.92	5.517
CIFAR10	$T=1$	$T=1$	0.95	11.36
	$T=3$	$T=3$	0.93	1.751
	$T=5$	$T=5$	0.90	10.62
LSUN	$T=1$	$T=1$	0.87	1.539
	$T=3$	$T=3$	0.89	4.341
	$T=5$	$T=5$	0.91	3.989
CelebA	$T=1$	$T=1$	0.91	11.65
	$T=3$	$T=3$	0.88	16.72
	$T=5$	$T=5$	0.96	14.11

从表 5～表 7 中可以发现，三个模型生成图像的质量排名顺序为 GP-WIRGAN＞WIRGAN＞GRAN.

（2）和非生成对抗网络模型的比较

该组对比实验在 MNIST 数据集上进行. 我们预先训练一个分类模型，其准确率达到 99.3%. 然后用 DAVE、DRAW、WIRGAN 和 GP-WIRGAN 生成 1000 张图像，将它们分别输入到预先训练好的分类模型中，得到分类错误率，错误率越低表明模型生成的图像质量越高. 分类结果如表 8 所示，忽略分

表 8 和非生成对抗网络模型的分类结果

model	Error/%
DAVE[24]	10.93
DRAW[6]	5.21
WIRGAN	1.63
GP-WIRGAN	0.99

类模型本身的影响，在同一个分类标准下，WIRGAN 和 GP-WIRGAN 的错误率要比 DAVE 和 DRAW 的低很多，并且 GP-WIRGAN 的错误率是最低的.

4.3.5　初始得分比较

初始得分评价的是生成图像的多样性. 表 9 是不同时间步骤下 WIRGAN 和 GP-WIRGAN 的初始得分. 从表中可以看到，两个模型的初始得分均是随着时间步骤的增大而增高，表明生成图像的多样性越高，这也说明循环结构有利于提高样本的多样性. 但是 $T=5$ 时的初始得分和 $T=3$ 时的初始得分相差不大. 值得注意的是相比于其他三个数据集，CIFAR10 的初始得分高很多，这是因为 CIFAR10 数据集中包含相对其他数据集更多类别，而且类别间的跨度更大，所以更容易得到不同的样本. 同时可以看到，GP-WIRGAN 的每一项初始得分都要比 WIRGAN 的每一项高.

表 9　不同时间步骤下 WIRGAN 和 GP-WIRGAN 的初始得分

Dataset	WIRGAN			GP-WIRGAN		
	$T=1$	$T=3$	$T=5$	$T=1$	$T=3$	$T=5$
MNIST	1.052±0.002	1.055±0.003	1.053±0.003	1.053±0.003	1.055±0.002	1.055±0.003
CIFAR10	7.73±0.03	7.77±0.09	7.77±0.06	7.76±0.03	7.83±0.06	7.84±0.02
LSUN	3.00±0.12	3.10±0.07	3.11±0.10	3.05±0.10	3.15±0.15	3.17±0.09
CelebA	1.92±0.20	2.00±0.20	2.00±0.10	2.10±0.10	2.12±0.09	2.15±0.10

此外，我们还对比了另一些无监督 GAN 模型如 ALI[25]，EGAN-Ent-VI[26]，DFM[27] 在 CIFAR 数据集上的初始得分，实验结果如图 7 所示，大部分无监督生成模型的初始得分都没有本文提出的模型的初始得分高，除 WGAN-GP ResNet 和 Progressive Growing of GAN 外. 进一步分析，WIRGAN 和 GP-WIRGAN 虽然用到了循环结构的生成模型，但是单个结构中神经网络层结构简单，而 WGAN-GP ResNet 是用 ResNet 构建的，Progressive Growing of GAN 虽然也是采用逐层增长的方式，但是其结构类似 ResNet，所以它们对数据的理解更深刻，生成的样本的多样性更高.

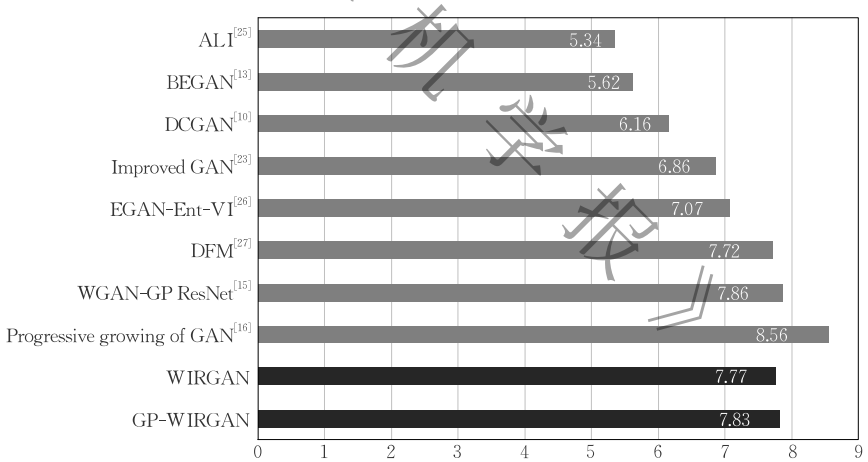

图 7　和其他 GAN 模型的初始得分对比结果

4.3.6　时间效率比较

虽然 WIGRAN 和 GP-WIGRAN 的生成器都是由多个结构相同的神经网络模型构成的，但是在神经网络的设计上存在区别，所以在计算速度上也存在差异. 我们统计了各个模型每迭代一次所需的计算时间，以 s 为单位. 如表 10 所示，在相同时

表 10　每迭代一次模型所需要的时间

Model	T	MNIST	CIFAR10	LSUN	CelebA
GRAN	1	0.0301±0.0001	0.1130±0.0002	0.2612±0.0002	0.3781±0.0002
GRAN	3	0.0302±0.0001	0.2015±0.0002	0.3455±0.0002	0.5355±0.0002
GRAN	5	0.0303±0.0001	0.3100±0.0002	0.4384±0.0002	0.7119±0.0002
WIRGAN	1	0.0074±0.0001	0.0700±0.0002	0.2498±0.0002	0.3348±0.0002
WIRGAN	3	0.0112±0.0001	0.0846±0.0002	0.3420±0.0002	0.5166±0.0002
WIRGAN	5	0.0116±0.0001	0.1424±0.0002	0.4108±0.0002	0.6516±0.0002
GP-WIRGAN	1	0.0073±0.0001	0.0626±0.0002	0.2422±0.0002	0.2676±0.0002
GP-WIRGAN	3	0.0107±0.0001	0.0817±0.0002	0.2894±0.0002	0.4168±0.0002
GP-WIRGAN	5	0.0115±0.0001	0.1262±0.0002	0.3521±0.0002	0.5535±0.0002

间步骤下,三个模型在每次迭代的计算时间关系:$T(\text{GRAN})>T(\text{WIRGAN})>T(\text{GP-WIRGAN})$,其中 GP-WIRGAN 计算用时最短.另外,从表 10 中可以看出,同一个模型迭代一次的所需的计算时间是随着时间步骤越大而递增的.但是 GP-WIRGAN 在相邻两个时间步骤之间的计算时间差波动幅度都要比 GRAN 和 WIRGAN 小.

4.4 可视化结果

4.4.1 WIRGAN 和 GP-WIRGAN 的生成结果

图 8、图 9 分别为 $T=3$ 时 WIRGAN 和 $T=5$ 时 WIRGAN 在 CelebA 和 LSUN-church 上的生成过程,图 10、图 11 分别为 $T=3$ 时 WIRGAN 和 $T=5$ 时 GP-WIRGAN 在 CelebA 和 LSUN 上的生成过程.从图中可以看到,相同时间步骤下,在每一时刻 GP-WIRGAN

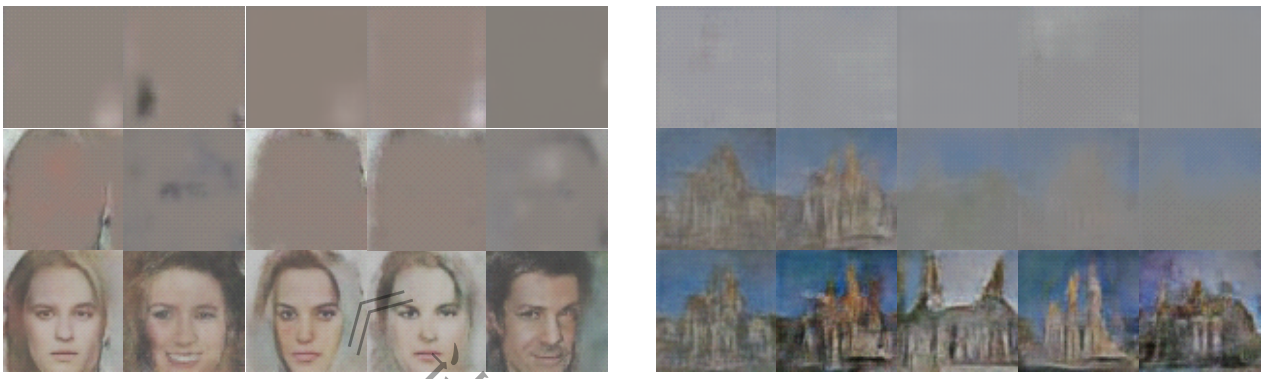

图 8 $T=3$ 时 WIRGAN 在 CelebA(左)和 LSUN-church(右)上的生成过程

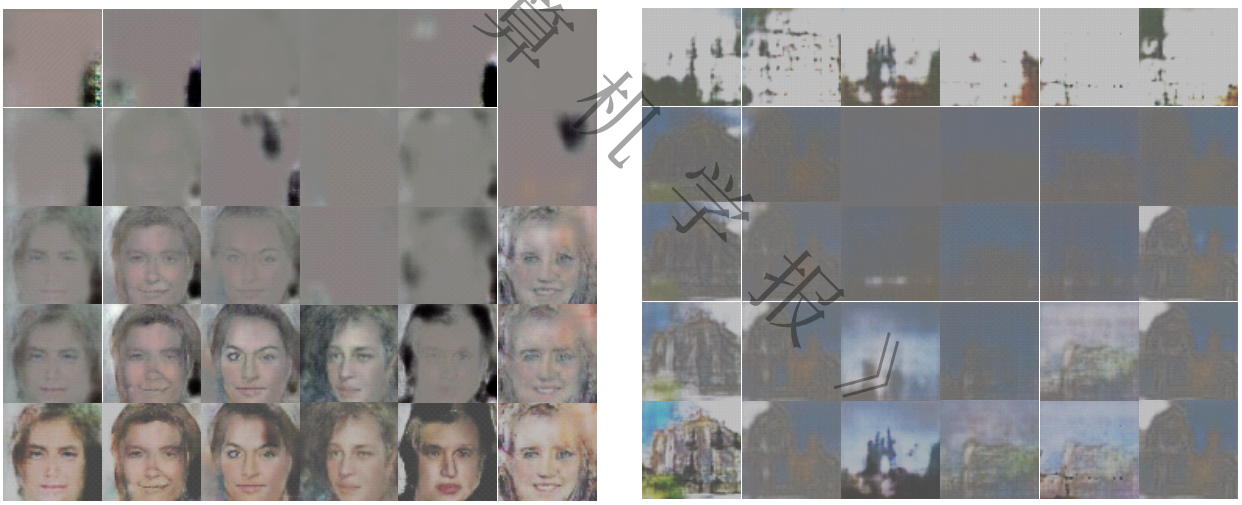

图 9 $T=5$ 时 WIRGAN 在 CelebA(左)和 LSUN-church(右)上的生成过程

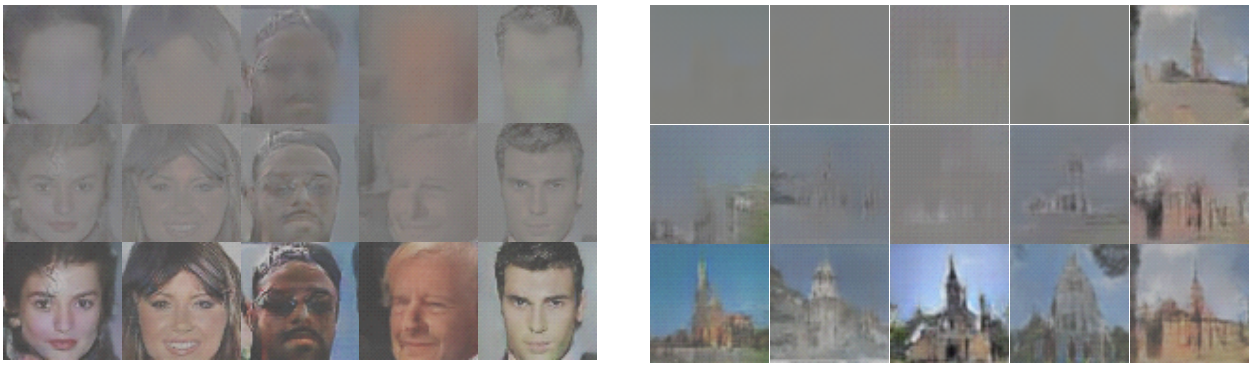

图 10 $T=3$ 时 GP-WIRGAN 在 CelebA(左)和 LSUN-church(右)上的生成过程

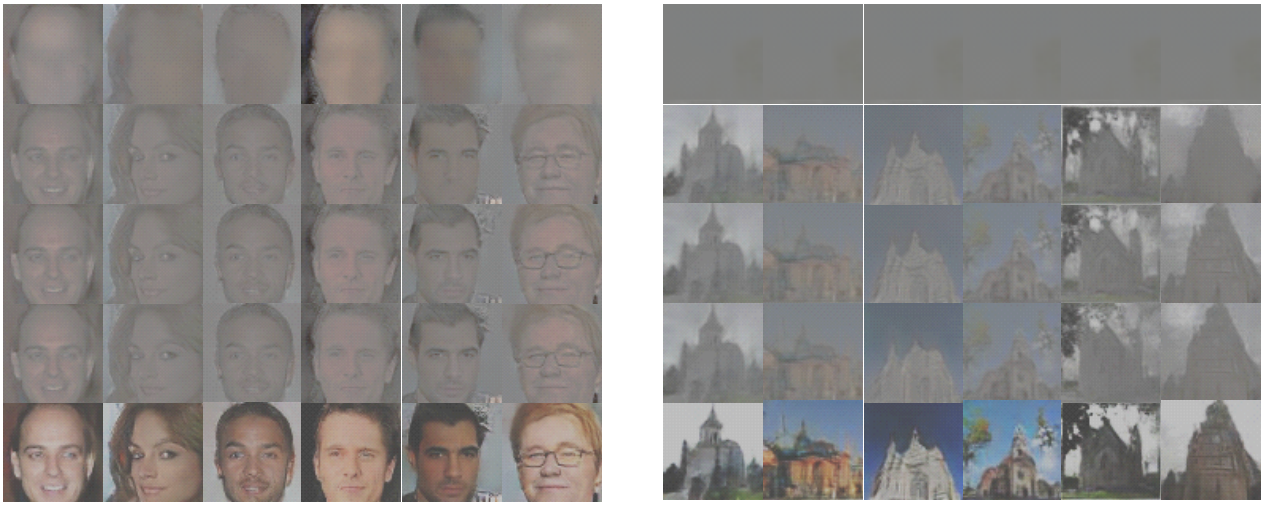

图 11　$T=5$ 时 GP-WIRGAN 在 CelebA(左)和 LSUN-church(右)上的生成过程

的生成结果均要比 WIRGAN 的生成结果清晰，颜色更丰富，线条更流畅. 总体而言，GP-WIRGAN 生成过程更流畅，生成的图像更锐利，且更具有解释性.

4.4.2　不同模型在相同数据集上的生成结果

图 12 中展示了三种不同模型的生成结果，从左到右依次为 $T=3$ 时的 GRAN、WIRGAN、GP-WIRGAN. GRAN 生成的数字不完整并且比较模糊，WIRGAN 和 GP-WIRGAN 生成的数字更加锐利，生成数字的多样性更高.

从图 13 中可以看出，GRAN 生成的样本缺乏多样性，图像不完整，缺乏可解释性；WIRGAN 的样本多样性相比于 GRAN 的更高，但是样本可解释性仍然不高；GP-WIRGAN 的生成结果比前两个模型的效果更好、图像质量更高且具有多样性.

图 12　$T=3$ 时 GRAN(左)、WIRGAN(中)、GP-WIRGAN(右)生成的数字

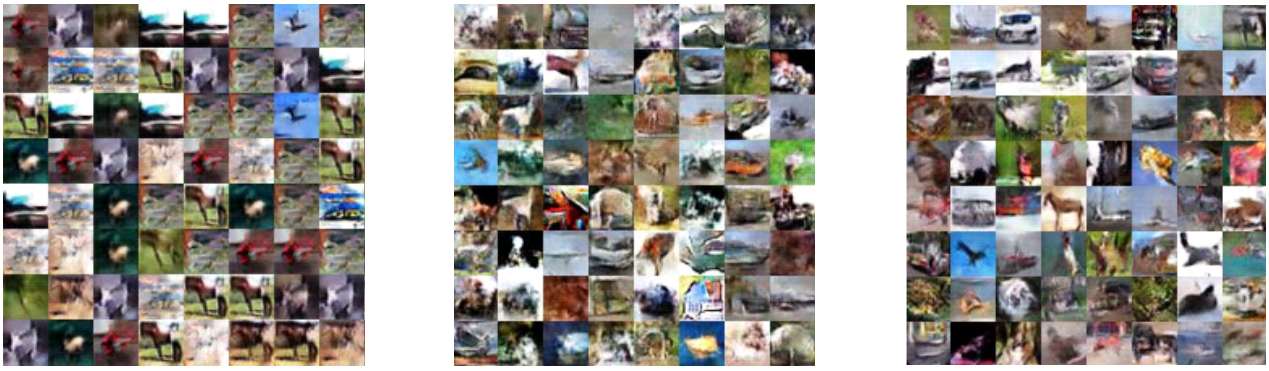

图 13　$T=3$ 时 GRAN(左)、WIRGAN(中)、GP-WIRGAN(右)生成的 CIFAR 样本

4.4.3　生成样本的质量和损失函数之间的关系

GP-WIRGAN 判别器的函数反映了生成图像的质量. 图 14 为 $T=3$ 时 GP-WIRGAN 生成样本和损失函数的关系图，可以看出，模型训练初期，生

成图像模糊且不可解释，随着损失函数的收敛，生成图像的图像信息逐渐丰富，可解释性变强.

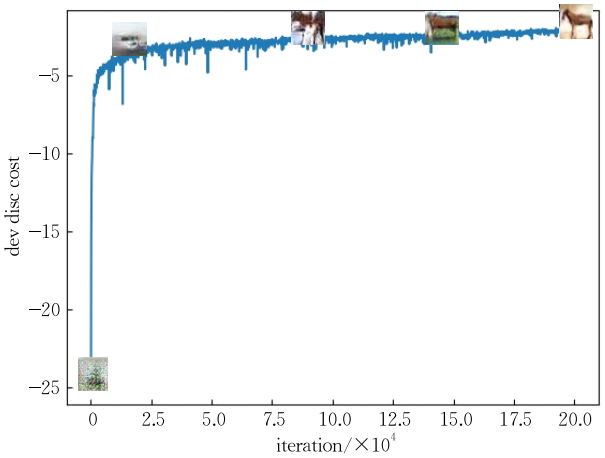

图 14 $T=3$ 时 GP-WIRGAN 生成样本和损失函数的关系

4.4.4 模型在 LSUN 上的生成结果

图 15 展示了 GP-WIRGAN 在 LSUN 上的生成样本.

图 15 GP-WIRGAN 生成的 LSUN-church 样本

5 结论与展望

本文主要对生成对抗网络模型进行研究，从新的距离度量角度对生成模型与判别模型进行改进以解决生成对抗网络训练难的问题，同时提出具有循环结构的生成模型. 结合上述两方面提出了 WIRGAN 模型，同时针对 WIRGAN 难以优化的问题，本文提出了进一步的改进模型 GP-WIRGAN. 从实验结果中可以发现本文提出的方法进一步提高了模型性能，使模型对数据具有一定理解，生成具有更高质量的样本图片. 但是，模型的神经网络结构和初始化方法简单，学习能力相对于其他优秀的神经网络还有差距，其生成结果的质量有待提高. 下一步准备结合 ResNet 等优秀的网络结构，采用更科学合理的初始化方法 Xavier，进一步改进模型，并利用模型强大的学习能力进行跨模态数据检索.

参 考 文 献

[1] Wang Kun-Feng, Gou Chao, Duan Yan-Jie, et al. Generative adversarial networks: The state of the art and beyond. Acta Automatica Sinica, 2017, 43(3): 321-332(in Chinese)
(王坤峰，苟超，段艳杰等. 生成式对抗网络 GAN 的研究进展与展望. 自动化学报，2017，43(3)：321-332)

[2] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets//Proceedings of the Advances in Neural Information Processing Systems. Montreal, Canada, 2014: 2672-2680

[3] Kingma D P, Welling M. Auto-encoding variational Bayes. arXiv eprint arXiv:1312.6114, 2013

[4] Vincent P, Larochelle H, Bengio Y, et al. Extracting and composing robust features with denoising autoencoders// Proceedings of the 25th International Conference on Machine Learning. Helsinki, Finland, 2008: 1096-1103

[5] Rezende D J, Mohamed S, Wierstra D. Stochastic backpropagation and approximate inference in deep latent Gaussian models//Proceedings of the International Conference on Machine Learning. Beijing, China, 2014, 2

[6] Gregor K, Danihelka I, Graves A, et al. DRAW: A recurrent neural network for image generation//Proceeding of the 32nd International Conference on Machine Learning. Lille, France, 2015: 1462-1471

[7] Ratliff L J, Burden S A, Sastry S S. Characterization and computation of local Nash equilibria in continuous games// Proceedings of the 51st Annual Allerton Conference on Communication, Control, and Computing. Monticello, USA, 2013: 917-924

[8] Mirza M, Osindero S. Conditional generative adversarial nets. arXiv eprint arXiv:1411.1784, 2014

[9] Denton E L, Chintala S, Fergus R. Deep generative image models using a Laplacian pyramid of adversarial networks// Proceedings of the Advances in Neural Information Processing Systems. Montreal, Canada, 2015: 1486-1494

[10] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv eprint arXiv:1511.06434, 2015

[11] Chen X, Duan Y, Houthooft R, et al. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets//Proceedings of the Advances in Neural Information Processing Systems. Barcelona, Spain, 2016: 2172-2180

[12] Im D J, Kim C D, Jiang H, et al. Generating images with recurrent adversarial networks. arXiv eprint arXiv: 1602.05110, 2016

[13] Zhao J, Mathieu M, LeCun Y. Energy-based generative adversarial network. arXiv eprint arXiv:1609.03126, 2016

[14] Arjovsky M, Chintala S, Bottou L. Wasserstein GAN. arXiv eprint arXiv:1701.07875, 2017

[15] Gulrajani I, Ahmed F, Arjovsky M, et al. Improved training of Wasserstein GANs//Proceedings of the Advances in Neural Information Processing Systems. California, USA, 2017: 5769-5779

[16] Karras T, Aila T, Laine S, et al. Progressive growing of GANs for improved quality, stability, and variation. arXiv eprint arXiv:1710.10196, 2017

[17] Peng Y, Qi J, Yuan Y. CM-GANs: Cross-modal generative adversarial networks for common representation learning. arXiv eprint arXiv:1710.05106, 2017

[18] Sriram A, Jun H, Gaur Y, et al. Robust speech recognition using generative adversarial networks. arXiv eprint arXiv: 1711.01567, 2017

[19] Deng L. The MNIST database of handwritten digit images for machine learning research. IEEE Signal Processing Magazine, 2012, 29(6): 141-142

[20] Recht B, Roelofs R, Schmidt L, et al. Do CIFAR-10 classifiers generalize to CIFAR-10?. arXiv eprint arXiv:1806.00451, 2018

[21] Yu F, Seff A, Zhang Y, et al. LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv eprint arXiv:1506.03365, 2015

[22] Liu Z, Luo P, Wang X, et al. Deep learning face attributes in the wild//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 3730-3738

[23] Salimans T, Goodfellow I, Zaremba W, et al. Improved techniques for training GANs//Proceedings of the Advances in Neural Information Processing Systems. Barcelona, Spain, 2016: 2226-2234

[24] Im D J, Ahn S, Memisevic R, et al. Denoising criterion for variational auto-encoding framework//Proceedings of the Association for the Advancement of Artificial Intelligence. San Francisco, USA, 2017: 2059-2065

[25] Dumoulin V, Belghazi I, Poole B, et al. Adversarially learned inference. arXiv eprint arXiv:1606.00704, 2016

[26] Dai Z, Almahairi A, Bachman P, et al. Calibrating energy-based generative adversarial networks. arXiv eprint arXiv: 1702.01691, 2017

[27] Warde-Farley D, Bengio Y. Improving generative adversarial networks with denoising feature matching//Proceedings of the 5th International Conference on Learning Representations. Toulon, France, 2017: 1-11

FENG Yong, Ph. D., professor. His current research interests include big data analysis and data mining, artificial intelligence and big data processing, deep learning and big data retrieval.

ZHANG Chun-Ping, M. S. His current research interests include deep learning and image retrieval.

QIANG Bao-Hua, Ph.D., professor. His current research interests include big data processing and information retrieval.

ZHANG Yi-Yang, M. S. His current research interests include deep learning and big data retrieval.

SHANG Jia-Xing, Ph. D., associate professor. His current research interests include artificial intelligence and big data processing.

Background

This work is a part of the "Image and Text Unified Retrieval based on Semantic Deep Understanding", which is mainly supported by the National Nature Science Foundation of China under Grant No. 61762025 and National Key R&D Program of China under Grant No. 2017YFB1402400. In the process of image retrieval, the user's retrieval intention may be ambiguous, or the user lacks understanding of semantic information of input image, which makes the image retrieval results difficult to understand and evaluate, and the retrieval quality is difficult to guarantee. We carry on image and text unified retrieval research based on semantic deep understanding, and the basic starting point of the research is to narrow the semantic gap between the image primitive semantics information and the human perception information. Aiming at the problem of object recognition and classification accuracy in complex multi-label images, we propose a novel training algorithm, which pre-trains deep neural network using single-label image and fine-tunes the deep neural network using multi-label image. At the same time, we can reduce the candidate boxes combining with abjectness detection technology. Aiming at the isomorphism problem of heterogeneous feature spaces of image and text, we propose heterogeneous space mapping and normalization algorithm to construct unified feature vector model based on deep canonical correlation analysis, and we use the normalized feature to tune deep neural network training process and optimize image and text feature extraction model. In order to improve the user's experience of image and text unified retrieval, a novel image and text semantic automatic summary algorithm is proposed. We consider retrieval time, image and text semantic relevance, user satisfaction and other factors, and a subjective and objective combined sorting and recommendation algorithm is proposed. Finally, we can narrow the semantic gap and achieve efficient and accurate semantics image and text unified retrieval.