doi:10.11835/j.issn.1000-582X.2020.03.006

一种基于自注意力机制的文本图像生成对抗网络

黄宏宇,谷子丰
(重庆大学 计算机学院,重庆 400044)

摘要:图像自动生成一直以来都是计算机视觉领域的一项重要挑战,其中的文本到图像的生成更是图像生成领域的重要分支。随着深度学习技术迅猛发展,生成对抗网络的出现使得图像生成领域焕发生机,借助生成对抗网络能够生成较为生动且多样的图像。本文将自注意力机制引入生成对抗网络,提出 GAN-SelfAtt 以提升生成图像的质量。同时,使用 WGAN、WGAN-GP 2 种生成对抗网络框架对 GAN-SelfAtt 进行实现。实验结果表明,自注意力机制的引入能够提高生成图像的清晰度,这归功于自注意力机制弥补了卷积运算中只能计算局部像素区域内的相关性的缺陷。除此之外,GAN-SelfAtt 在训练时有着更好的稳定性,避免了原始生成对抗网络中的模式坍塌问题。

关键词:文本生成图像;生成对抗网络;自注意力机制;深度学习
中图分类号:TP311 **文献标志码**:A **文章编号**:1000-582X(2020)03-055-07

A generative adversarial network based on self-attention mechanism for text-to-image generation

HUANG Honggu, *GU Zifeng*
(College of Computer Science, Chongqing University, Chongqing 400044, P. R. China)

Abstract: Automatic image generation is a challenging problem in computer vision for a long time. As a branch of this area, there are also challenges in text-to-image generation. With the fast development of deep learning, generative adversarial networks (GANs) give a new inspiration to the image generation because it can generate highly compelling images of various categories. In this paper, we introduce the self-attention mechanism to GAN and propose GAN-SelfAtt to improve the quality of images. Meanwhile, we implement GAN-SelfAtt using two different GAN frameworks, i.e., WGAN and WGAN-GP. The experimental results show that self-attention mechanism improves the resolution of generated images. The reason of this improvement is that the self-attention mechanism fixes the defect of convolution computation which only calculates the correlation in the local pixel region. In addition, our results show that the stability of GAN-SelfAtt during the training process is improved. This fixes the problem of mode collapse which appears in the original GANs.

Keywords: text-to-image generation; generative adversarial networks; self-attention mechanism; deep learning

---

收稿日期:2019-03-27
基金项目:重庆市自然科学基金资助项目(cstc2014jcyjA40030)。
Supported by Chongqing Research Program of Basic Research and Frontier Technology (cstc2014jcyjA40030).
作者简介:黄宏宇(1979—),男,博士,副教授,主要从事物联网,隐私保护,深度学习方向研究,(E-mail)hyhuang@cqu.edu.cn。

近年来,深度学习技术在计算机视觉领域发展迅猛,一些计算机视觉中的基本问题,诸如图像识别与分类、图像的文本描述以及图像分割等问题都取得了突破性进展。文本生成图像依赖于图像生成中的模型。相较于对图像进行识别与分类的判别模型,生成模型往往有着较高的计算复杂度,因而其研究进展也比较缓慢。早期生成模型往往基于概率图模型,涉及最大似然估计、马尔科夫链、近似法等方法。其中变分自编码器(VAE, variational auto-encoder)[1]采用近似法,将生成图像的建模问题转换为概率图模型来解决问题,模型效果受限于所假设的近似分布好坏。PixelRNN 方法[2]则通过为对像素序列的条件分布进行预测来进行图像生成,并假设每个像素的取值只依赖于像素序列中位于之前的像素取值。而上述方法及后续改进只适用于生成结构简单、主体单一的图像,诸如 MNIST 和 CIFAR-10 等数据集中的图像,因为这些方法对于真实世界的数据需要大量的先验知识,同时方法本身也需要庞大的计算量。于是 Goodfellow 等[3]提出了生成对抗网络(GAN,generative adversarial network),其简单而有效的特点在图像生成等领域受到了广泛关注。

生成对抗网络通常基于神经网络,由生成器和判别器 2 个神经网络构成。在训练优化的过程中,判别器的目标是尽可能最优地判断输入图像是否来自于真实图像分布,即是否是真实图像;生成器的目标则是生成更加逼近真实图像分布的伪图像,即尽可能地欺骗判别器,将其伪图像判定为来自真实图像分布。当生成器与判别器在对抗训练过程中达到纳什均衡时,认定模型中的生成器具备了生成较为真实图像的能力。原始生成对抗网络由于其生成内容不可控制,Mirza 等[4]提出了基于条件生成对抗网络(CGAN,conditional gan),将条件变量作为附加信息输入生成器中以约束生成过程。同时生成对抗网络存在模式坍塌(mode collapse)问题[13],即生成图像的模式集中于少数几个而缺乏多样性。后续研究提出了 Deep Convolutional GAN (DCGAN)[5]、Wasserstein GAN (WGAN)[6]、Wasserstein GAN-Gradient Penalty (WGAN-GP)[7]等方法解决模式坍塌问题,其中 DCGAN 从模型架构上入手改进,而 WGAN 和 WGAN-GP 则通过改进真实分布与模型分布之间的评估标准解决问题,均取得了较为优秀的结果。

文本生成图像的研究通常需要将文本内容转换为向量以作为附加信息,约束训练过程中图像的生成,这一阶段通常采用文本描述向量化技术[8]来完成。文献[9]提出 GAN-CLS 架构将文本嵌套与生成对抗网络相结合通过文本生成图像。GAN-CLS 使用 DCGAN 框架来实现文本到图像的生成,但其结果清晰度较低且缺乏多样性。研究提出了文本到图像的生成对抗网络 GAN-SelfAtt,将自注意力机制与 WGAN、WGAN-GP2 种生成对抗网络架构相结合,达到图像生成效果提升。实验结果表明,自注意力机制的引入使得模型 GAN-SelfAtt 相对于之前的方法 GAN-CLS 明显提高了生成图像的清晰度,同时还提高了生成图像的多样性。

1 文本到图像生成对抗网络的构建

1.1 生成对抗网络

文本生成图像的一项挑战在于基于文本描述的图像分布是高度模态化的,基于条件的多模态正是生成对抗网络最擅长的应用方式。在生成对抗网络中,生成器 G 为反卷积神经网络,判别器 D 为卷积神经网络。G 的原始输入,z 为服从正态分布的随机噪音,即 $z\sim N(0,1)$,输出为伪图像 $G(z)$;D 的输入分别为来自真实数据集的图像,x 与生成器生成的图像 $G(z)$,输出可以看作判别器给出的这张图像真实度的分值 $D(x)$与 $D(G(z))$,取值在 0 到 1 之间。生成对抗网络的训练过程是生成器与判别器对抗博弈的过程,其中判别器的目标是给图像标注真实度分数,尽可能给较为真实的图像标注高分而来自生成器生成的伪图像尽量标注低分;生成器的目标则是生成更加真实的图像,以混淆判别器的判断,尽可能让判别器为其生成的图像标注高分,即让判别器认定其更可能来自于真实的数据分布。根据这样对抗训练的思想,生成对抗网络的损失函数可以用以下公式表示

$$\min_G \max_D V(G,D)=E_{x\sim p_x}[\log D(x)]+E_{z\sim p_z}[\log[1-D(G(z))]], \tag{1}$$

其中:P_x 为真实图像的数据分布;P_z 为输入噪声的数据分布。训练过程中,判别器与生成器并不是同时训练的。训练判别器时,生成器参数固定,训练过程只更新判别器中的参数,训练生成器时同理。因此,判别器与生成器损失函数分别定义如下

$$L_D=-E_{x\sim p_x}[\log D(x)]+E_{z\sim p_z}[\log[1-D(G(z))]], \tag{2}$$

$$L_G = -E_{z\sim pz}[\log[1-D(G(z))]], \tag{3}$$

其中，L_D 与 L_G 为原始 GAN 的目标函数。但是原始 GAN 存在模式坍塌问题，样本模式单一缺乏多样性。后续的研究中，WGAN 从根本原因出发改进这一问题。WGAN 的设计者指出原始 GAN 使用交叉熵损失即 JS 散度衡量真实分布与模型生成图像分布之间的差异，当两者分布在训练初期不存在相交时，采用 JS 散度的训练结果较差，不能够为生成器的训练提供有效的梯度从而导致模式坍塌。而 WGAN 采用 Wasserstein 距离衡量真实分布与模型分布之间的差异，从根本上解决了模式坍塌的问题。WGAN 的损失函数由如下公式定义

$$\min_{G}\max_{D\in L} V(G,D) = E_{x\sim p_x}[D(x)] + E_{z\sim p_z}[1-D(G(z))], \tag{4}$$

式中，$D\in L$，L 是 $1-L$ipschitz 函数，为了满足 $1-L$ipschitz 条件，需要对判别器中的权重加以约束，WGAN 中采用权重裁剪的方法以满足约束。最终 WGAN 从根本上解决原始 GAN 中的问题，但其进行权重裁剪的方法过于直接，在特定情况下会产生梯度消失于梯度爆炸等问题导致训练困难，因此后续提出了 WGAN-GP 方法进行了改进。WGAN-GP 舍弃了权重裁剪的方法，使用梯度惩罚来约束判别器权重，其判别器的损失函数变更为

$$L_D = -E_{x\sim p_x}[D(x)] + E_{z\sim p_z}[1-D(G(z))] + E_{c\sim P_c}[(\| \nabla_c D(c)\|_2 - 1)^2], \tag{5}$$

其中，c 为真实图像 x 与伪图像 $G(z)$ 连线上的随机差值采样。最终，WGAN-GP 同样从根本原因出发，解决了原始 GAN 中训练困难导致模式坍塌的问题，在后续的研究中广泛应用。

1.2　文本描述的向量化

为了在训练模型的过程中引入文本描述与图像之间的关联，借鉴 Reed 等人[8]提出的方法，使用深度卷积网络与递归神经网络学习文本描述与图像的对应函数，首先文本分类器通过最小结构化损失函数获得

$$\frac{1}{N}\sum_{i=1}^{N}\Delta(y_i, f_v(v_n)) + \Delta(y_i, f_t(t_i)), \tag{6}$$

其中：$\{(v_i, t_i, y_i): i=1,\cdots,N\}$ 为训练数据集；Δ 为 0－1 损失；v_i 为图像样本；t_i 为相应的文本描述；y_i 则为类别标签。分类器 f_v 与 f_t 的参数化表示如下

$$f_v(v) = \underset{y\in Y}{\operatorname{argmax}} E_{t\sim T(y)}[\boldsymbol{\varphi}(v)^{\mathrm{T}}\boldsymbol{\varphi}(t)], \tag{7}$$

$$f_t(v) = \underset{y\in Y}{\operatorname{argmax}} E_{v\sim V(y)}[\boldsymbol{\varphi}(v)^{\mathrm{T}}\boldsymbol{\varphi}(t)], \tag{8}$$

式中：$\boldsymbol{\varphi}(v)$为图像编码器，一般为一个深度卷积神经网络；$\boldsymbol{\varphi}(t)$则是字符级的文本编码器，一般是字符级别的 CNN 或者 LSTM 结构。$T(y)$代表类别为 y 的文本描述的集合，$V(y)$代表类别为 y 的图像样本的集合。参照 Akata 等人[11]提出的方法，需将上述模型结果梯度反向传播至 $\boldsymbol{\varphi}(t)$中得到一个有区分能力的文本编码器，工作采取 CNN 与 LSTM 相结合的编码器架构。

2　基于自注意力机制的 GAN-SelfAtt

2.1　模型框架

当前文本到图像的生成研究大部分借助生成对抗网络架构。例如最早被提出的 GAN-CLS 能够初步生成与文本描述匹配的图像，但其架构基于 DCGAN，在生成架构中仅使用多层卷积层难易生成图像清晰度较高的图像。因此，以提高生成图像的清晰度为目标，提出基于自注意力机制的文本到图像生成对抗网络 GAN-SelfAtt。

笔者提出的文本到图像生成对抗网络 GAN-SelfAtt 结合了自注意力机制以提高生成图像的质量，其中生成器的架构如图 1 所示。生成器为反卷积神经网络，判别器为卷积神经网络，并且判别器在相同位置引入了自注意力机制。生成器的输入为服从正态分布的随机噪音向量 $\boldsymbol{z}$ 与文本描述 t 的文本描述向量 $\boldsymbol{\varphi}(t)$的结合，输出为生成的伪图像 $G(\boldsymbol{z}, \boldsymbol{\varphi}(t))$，其中 $\boldsymbol{z}\sim N(0,1)$。判别器的作用不仅要判别图像是否真实，同时还要判断文本内容与图像是否匹配。因此，判别器的输入为来自真实数据集的图像 x 与生成的伪图像 $G(\boldsymbol{z},$

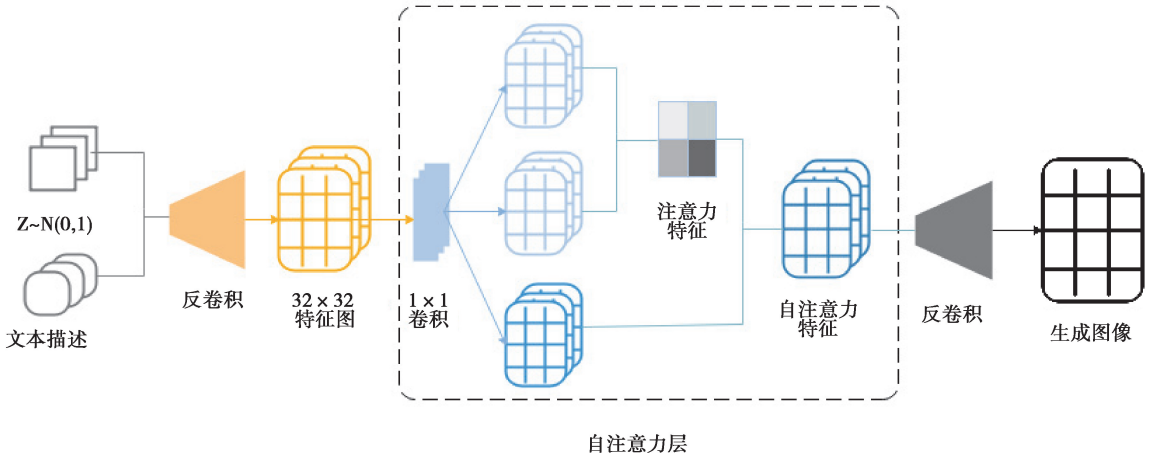

图1 基于自注意力机制的文本到图像生成对抗网络 GAN-SelfAtt

Fig. 1 Text-to-Image GAN with Self-Attention mechanism, GAN-SelfAtt

$\boldsymbol{\varphi}(t)$),同时在判别器网络的中间部分辅以相应的文本向量 $\boldsymbol{\varphi}(t)$,以此判别图像是否与文本内容相符。由此得模型训练的对抗损失函数如下

$$L_D = -E_{x\sim P_{data},t\sim P_{data}}[\min(0,-1+D(x,\boldsymbol{\varphi}(t)))] - E_{z\sim p_z,t\sim P_{data}}[\min(0,-1-D(G(\boldsymbol{z},\boldsymbol{\varphi}(t))))], \tag{9}$$

$$L_G = -E_{z\sim P_z,t\sim P_{data}} D(G(\boldsymbol{z},\boldsymbol{\varphi}(t)))。 \tag{10}$$

GAN-SelfAtt 的生成器与判别器轮番训练,训练生成器时判别器参数固定,训练判别器时生成器参数固定。模型使用 Leaky ReLU 作为激活函数。框架优化方法的选择对于 DCGAN 与 WGAN-GP 的实现选取 Adam[14]优化方法,WGAN 的实现尽量不使用基于动量的优化方法,因此选取 RMSprop[15]优化方法。

2.2 自注意力机制

大多数进行图像生成的生成对抗网络都依赖卷积神经网络中的卷积层构建网络。经过多个卷积层后,模型处理图像的长距离像素区域之间相关性(long-range dependencies in pixels)计算效率较低,不能考虑长距离像素区域之间的相关性。借助 Goodfellow 等[10]提出基于自注意力机制的生成对抗网络(SAGAN,self-attention gan)中的自注意力机制。架构在生成对抗网络的卷积框架中引入了非局部模块,即自注意力模块,在原有的卷积层中间计算局部像素区域之间的注意力值,以弥补模型对于长相关性与局部相关性计算的不足。

自注意力模块位于卷积层中部,输入为上一层卷积层输出的特征图,输出则与原特征图相加得到下一层输入。来自上一隐层的特征图 $x\in R^{C\times N}$ 首先被转换到 2 个特定的特征空间 $\boldsymbol{f}$ 与 g 中以计算注意力值,其中 $\boldsymbol{f}(x)=\boldsymbol{W}_f x$,$g(x)=\boldsymbol{W}_g x$,

$$\beta_{j,i} = \frac{\exp(s_{ij})}{\sum_{i=1}^{N}\exp(s_{ij})},\text{且 } s_{ij} = \boldsymbol{f}(x_i)^{\mathrm{T}} g(x_j), \tag{11}$$

$$o_j = \sum_{i=1}^{N}\beta_{j,i} h(x_i),\text{且 } h(x_i) = \boldsymbol{W}_h x_i, \tag{12}$$

其值代表生成区域第 j 个区域时,模型关注第 i 个区域的注意程度。注意力模块的输出为 $o=(o_1,o_2,o_3,\cdots,o_N)\in R^{C\times N}$。最后,将注意力模块的输出加权与原特征图相加得到最终结果,以此作为下一隐层的输入。

$$y_i = \gamma o_i + x_i, \tag{13}$$

其中:y 为自注意力层的最终输出;γ 初始化为 0。上述计算中,$\boldsymbol{W}_g\in R^{\bar{C}\times C}$,$\boldsymbol{W}_f\in R^{\bar{C}\times C}$,$\boldsymbol{W}_h\in R^{C\times C}$ 均为可学习的权重矩阵,通过 1 X 1 的卷积运算实现,其中 $\bar{C}=\frac{C}{8}$。对于整个基于自注意力机制的生成对抗网络,同样是对生成器与判别器的损失函数进行交替训练,通过训练如下对抗损失函数以优化整个网络

$$L_G = -E_{z\sim P_z} D(G(z)), \tag{14}$$

$$L_D = -E_{x\sim Px}[\min(0, -1 + D(x))] - E_{z\sim P_z}[\min(0, -1 - D(G(z)))]。 \tag{15}$$

3　实验与结果

3.1　实验数据集与参数设置

实验部分，选取 Oxford-102 花朵数据集[12]，共包含 8189 张花朵图片，包含 102 个不同种类。输入图像的尺寸为 64×64×3，嵌套后的文本描述向量尺寸为 1024 维。在生成器和判别器中使用线性映射至 128 维以方便链接后续的卷积层。实验结果依次使用 WGAN、WGAN-GP 2 种生成对抗网络框架实现 GAN-SelfAtt 并与文本到图像生成的原始方案 GAN-CLS 对比结果。其中 GAN-SelfAtt 的 WGAN-GP 实现使用 Adam 优化器训练，判别器与生成器使用不同学习率，其中判别器学习率设置为 0.000 4，生成器学习率为 0.000 1，超参数 beta1 和 beta2 分别取值 0.0 和 0.9。GAN-SelfAtt 的 WGAN 实现则使用 RMSprop 优化器训练，判别器与生成器学习率统一，均设置为 0.000 2 以取得较优的结果。实验选取显卡 Nvidia GTX1070ti 于 windows10 平台进行各类生成对抗模型的训练。

3.2　实验结果与分析

实验阶段对原始的文本到图像生成方案 GAN-CLS 进行文本到图像的生成进行了复现，同时实现了提出的基于自注意力机制 GAN-SelfAtt 的 WGAN 和 WGAN-GP 实现，并与 GAN-CLS 进行对比。对于 GAN-SelfAtt 的 2 种实现，分别命名为 WGAN-SelfAtt 与 WGAN-GP-SelfAtt 以方便区分。图 2 展示了实现 3 种生成对抗网络在 Oxford-102 花朵数据集上的生成效果，为了使得结果更为直观，分别筛选出 3 种模型结果中生成质量较优的图像进行对比。由于生成对抗网络结果的评估标准较少，首先使用人工观察鉴别方式来评估模型的生成效果。可以看出最先提出的 GAN-CLS 方法虽然能生成较完整的图像，但图像细节处理不足，清晰度较低。这是由于 GAN-CLS 基于 DCGAN 的框架实现，在生成图像时仅有卷积层的操作不能获得图像区域之间的依赖，导致无法还原图像细节。而 GAN-SelfAtt 方法能够更好的还原图像细节，这归功于自注意力机制的引入。但是由于 WGAN-SelfAtt 基于框架 WGAN，使用权重裁剪方法来满足 Lipschitz 约束的方法太过勉强，导致梯度过于集中在权值裁剪的边缘，个别图像效果较差。而基于 WGAN-GP 实现的 WGAN-GP-SelfAtt 同样能够生成清晰度更高的图像，同时使用梯度惩罚的方法满足 Lipschitz 约束，保证了训练过程稳定，从而其整体生成图像的效果较好。

图 2　GAN-CLS、WGAN-SelfAtt、WGAN-GP-SelfAtt 生成图像效果对比

Fig. 2　Comparison of generated flowers using GAN-CLS, WGAN-SelfAtt, and WGAN-GP-SelfAtt.

上述结果直观对比了 3 种框架下生成图像的效果。研究还对生成图像与文本描述的匹配程度以及不同框架下生成图像的多样性进行了对比。图 3 的结果列出了 3 种框架下部分文本描述生成的相应图像。可看出 3 种框架均能生成与文本描述相匹配的图像。但是 3 种框架的不同之处在于 GAN-CLS 框架生成的图像

较为单一，而 WGAN-SelfAtt 与 WGAN-GP-SelfAtt 方法对同一种描述能够生成多种类别的图像。可以看出 WGAN-SelfAtt 与 WGAN-GP-SelfAtt 生成的图像具有多样性，这正是因为上述 2 种方法使用 Wasserstein 距离来衡量真实数据分布与模型分布之间的距离，从根本原因上解决原始 GAN 中的模式坍塌问题。而基于 DCGAN 的 GAN-CLS 是在神经网络架构上进行改进，使用 JS 散度来衡量模型分布与真实数据分布之间的距离，没能从根本上解决模式坍塌问题，因此对于部分结果同样不具有多样性且生成图像效果较差（如图 3 所示）。

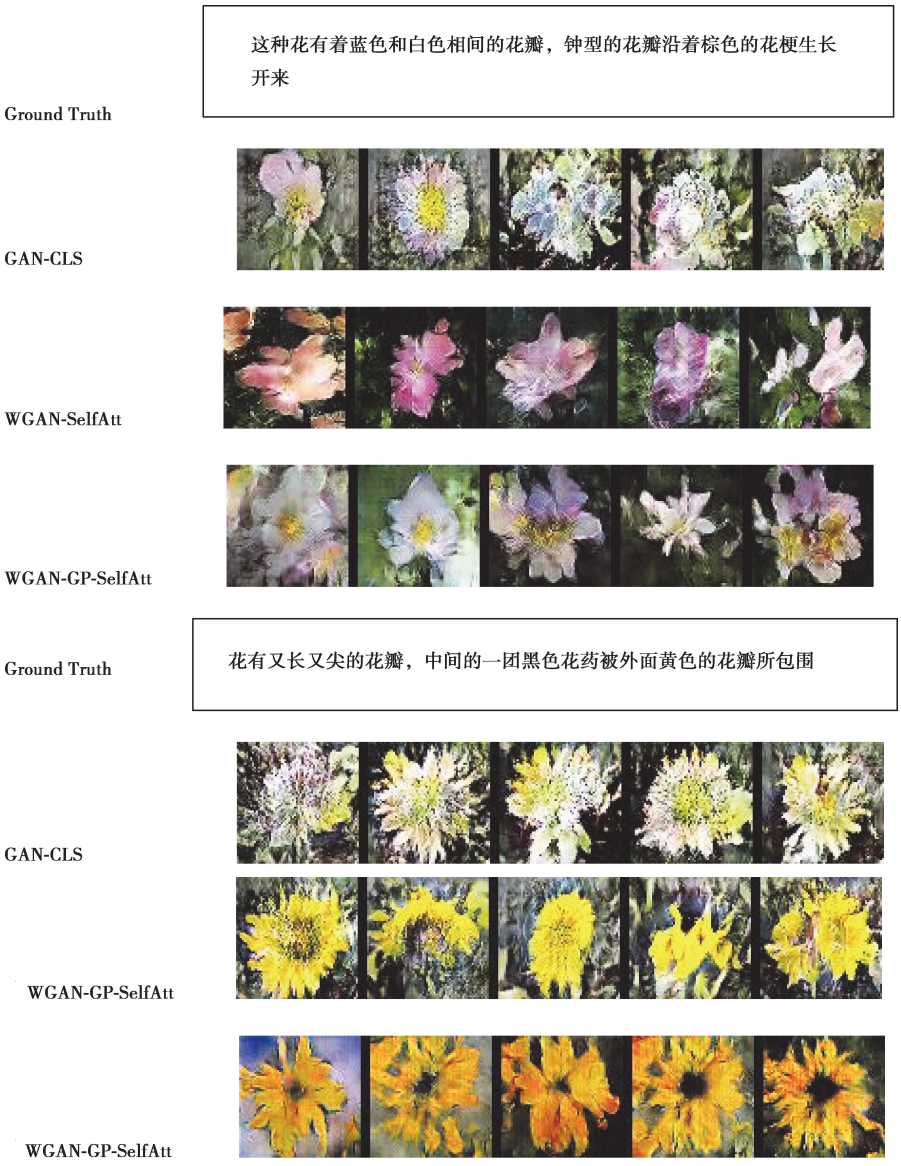

图 3　基于真实图像文本描述，GAN-CLS、WGAN-SelfAtt、WGAN-GP-Self-Att 的生成图像结果对比

Fig. 3　Based on the ground truth text description, the comparison of generated images from GAN-CLS, WGAN-SelfAtt, and WGAN-GP-SelfAtt respectively.

综合图 2 与图 3 中的结果，可以验证工作提出的基于自注意力机制的文本到图像的生成对抗网络 GAN-SelfAtt 在一定程度上提高了生成图像的清晰度，使得生成的图像结果更为逼真。除此之外，在实验中使用了 WGAN 与 WGAN-GP2 种生成对抗网络框架实现 GAN-SelfAtt。通过实验结果对比得出，基于 DCGAN 的文本到图像生成的解决方案 GAN-CLS 依旧会发生原始 GAN 中固有的模式坍塌问题，即部分生成结果效果较差，且多样性不足，生成图像结果较为单一；而基于 WGAN 与其改进方案 WGAN-GP 的 GAN-SelfAtt 能够从根本克服模式坍塌问题，生成的图像结果质量较高，且具有一定多样性。

4 结　论

实现了基于自注意力机制的文本到图像的生成对抗网络 GAN-SelfAtt，提升了生成图像的清晰度。同时还使用了 2 种生成对抗网络 WGAN 与 WGAN-GP 对 GAN-SelfAtt 进行实现，并与 GAN-CLS 方案进行了效果的对比。其中基于 DCGAN 实现文本到图像生成的原始方法 GAN-CLS 清晰度不高，缺乏多样性。WGAN 与 WGAN-GP 的引入解决了生成效果多样性不足的问题，同时借助 GAN-SelfAtt 中的自注意力机制提高了生成图像的清晰度。其中，生成对抗网络中的自注意力机制在原本的卷积层中间计算了不同局部像素区域之间的关系，主要解决了卷积操作只能处理局部像素间相关性，而不能计算局部区域之间长距离、多层的相关性缺陷，从而借助自注意力机制生成的图像质量更高。除此之外，当下由多个生成器与判别器组合的生成对抗网络框架是研究趋势，因此在未来的研究中，将致力于引入多个生成器与判别器的生成对抗网络以提高生成图像的质量与效果。

参考文献：

[1] Kingma D P, Welling M. Auto-encoding variational bayes[C/OL]. 2nd International Conference on Learning Representations (ICLR2014), NY: arXiv.org. 2014[2019-09-25]. https://dare.uva.nl/search? identifier=cf65ba0f-d88f-4a49-8ebd-3a7fce86edd7.

[2] Oord A, Kalchbrenner N, Kavukcuoglu K. Pixel recurrent neural networks[J/OL]. arXiv: Computer Vision and Pattern Recognition, 2016[2019-09-25]. https://arxiv.org/abs/1601.06759.

[3] Ian J. Goodfellow. Generative adversarial nets[C]// NIPS'14 Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2. Cambridge, MA, USA: MIT Press.2014:2672-2680.

[4] Mirza M, Osindero S. Conditional generative adversarial nets[J/OL]. arXiv: Learning, 2014[2019-09-25]. http://www.oalib.com/paper/4066323#.XfimLbNuJZQ.

[5] Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversaria networks[J/OL]. arXiv: Machine Learning, 2015[2019-09-25]. https://arxiv.org/abs/1511.06434.

[6] Arjovsky M, Chintala S, Bottou L. Wasserstein gan[J/OL]. arXiv: Machine Learning, 2017[2019-09-25]. https://arxiv.org/abs/1701.07875.

[7] Gulrajani I, Ahmed F, Arjovsky M, et al. Improved training of wasserstein gans[J/OL]. Computer Science, 2017[2019-09-25]. https://arxiv.org/abs/1704.00028.

[8] Reed S, Akata Z, Lee H, et al. Learning deep representations of fine-grained visual descriptions[C/OL]. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). NewYork, USA: IEEE, 2016(2016-12-12)[2019-09-25]. https://ieeexplore.ieee.org/document/7780382.

[9] Reed S, Akata Z, Yan X C, et al. Generative adversarial text to image synthesis[J/OL]. Neural and Evolutionary Computing, 2016[2019-09-25]. http://export.arxiv.org/abs/1605.05396.

[10] Zhang H, Goodfellow L, Metaxas D N, et al. Self-attention generative adversarial networks[J/OL]. arXiv: Machine Learning, 2018[2019-09-25]. https://www.researchgate.net/publication/325311774_Self-Attention_Generative_Adversarial_Networks.

[11] Akata Z, Reed S, Walter D, et al. Evaluation of output embeddings for fine-grained image classification[C/OL]. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). New York, USA: IEEE, 2015(2015-10-15)[2019-09-25]. https://ieeexplore.ieee.org/document/7298911.

[12] Nilsback M E, Zisserman A. Automated flower classification over a large number of classes[C/OL]. 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing. New York, USA: IEEE, 2008[2019-09-25]. https://ieeexplore.ieee.org/document/4756141.

[13] Salimans T, Goodfellow I, Zaremba W, et al. Improved techniques for training GANs[J/OL]. Conference on Neural Information Processing Systems, 2016[2019-09-25]. https://wenku.baidu.com/view/7d0398b6f80f76c66137ee06eff9aef8951e484a.html.

[14] Diederik P K, Jimmy B. Adam: a method for stochastic optimization[J/OL]. Computer Science.2014[2019-09-25]. https://www.researchgate.net/publication/269935079_Adam_A_Method_for_Stochastic_Optimization.

（编辑　侯　湘）