文章编号:1672-3961(2024)06-0049-08 DOI:10.6040/j.issn.1672-3961.0.2023.113

基于空间注意力及条件增强的文本生成图像方法

马军[1,2],车进[1,2*],贺愉婷[1,2],马鹏森[1,2]
(1.宁夏大学电子与电气工程学院, 宁夏 银川 750021; 2.宁夏沙漠信息智能感知重点实验室, 宁夏 银川 750021)

摘要:针对文本生成图像语义不一致、训练不稳定、生成图像单一等问题,在一种简单有效的文本生成图像基准模型上提出基于空间注意力及条件增强的文本生成图像模型。为提高训练过程的稳定性、增加生成图像的多样性,在原有模型基础上增加条件增强模型;从文本分布出发拟合图像分布,增加视觉特征的多样性,扩大表现空间,在原有的 DF-Block 模块中增加一层 Affine 仿射块。在判别器中加入空间注意力模型,提高文本与合成图像的语义一致性。试验结果表明,在 CUB 和 Oxford-102 数据集上,初始得分分别提高了 2.05%和 2.63%;在 CUB 和 COCO 数据集上,特征空间距离分别降低了 20.73%和 9.25%。本研究提出的模型生成的图像更具多样性且更接近真实图像。

关键词:文本生成图像;DF-GAN;条件增强模型;Affine 仿射块;空间注意力模型
中图分类号:TP391 **文献标志码**:A
引用格式:马军,车进,贺愉婷,等. 基于空间注意力及条件增强的文本生成图像方法[J].山东大学学报(工学版),2024,54(6):49-56.
MA Jun, CHE Jin, HE Yuting, et al. Text-to-image synthesis method based on spatial attention and conditional augmentation[J]. Journal of Shandong University (Engineering Science), 2024, 54(6):49-56.

Text-to-image synthesis method based on spatial attention and conditional augmentation

MA Jun[1,2], CHE Jin[1,2*], HE Yuting[1,2], MA Pengsen[1,2]
(1. School of Electronic and Electrical Engineering, Ningxia University, Yinchuan 750021, Ningxia, China; 2. Key Laboratory of Intelligent Sensing for Desert Information, Yinchuan 750021, Ningxia, China)

Abstract: For the problems such as inconsistent semantics of text-to-images, unstable training, and single generated images, a text-to-images model based on spatial attention and conditional augmentation was proposed on a simple and effective text-to-images benchmark model. To improve the stability of the training process and increase the diversity of generated images, a conditional augmentation model was added on the basis of the original model; starting from the text distribution to fit the image distribution, increasing the diversity of visual features and expanding the performance space, and adding an Affine block in the original DF-Block module. A spatial attention model was added to the discriminator to improve the semantic consistency of the text and the synthetic image. The experimental results showed that on the CUB and Oxford-102 datasets, inception score increased by 2.05% and 2.63% respectively; and on the CUB and COCO datasets, Fréchrt inception distance decreased by 20.73% and 9.25% respectively. The results proved that the images generated by the proposed model were more diverse and closer to real images.
Keywords: text-to-images; DF-GAN; conditional augmentation model; Affine block; spatial attention model

0 引言

近年来,生成对抗网络(generative adversarial network, GAN)在图像修复、图像风格迁移、增强超分辨率和文本图像生成等[1]方面广泛应用。其中,文本图像生成是 GAN 最重要的研究领域之一。构建一个由文本生成图像的模型,能够为大众提供极

收稿日期:2023-05-30
基金项目:国家自然科学基金资助项目(61861037);宁夏大学研究生创新研究基金资助项目(CXXM202223)
第一作者简介:马军(1996—),男,宁夏吴忠人,硕士研究生,主要研究方向为计算机视觉、图像生成及深度学习。
E-mail:1229012138@qq.com
*通信作者简介:车进(1973—),男,宁夏银川人,教授,硕士生导师,博士,主要研究方向为智能信息处理与模式识别以及多模态智能。
E-mail:koalache@126.com

大的便利[2]。

虽然文本生成图像的相关技术有了显著的发展,但是结合计算机视觉和自然语言处理[3]两大领域的任务一直都很具有挑战性[4]。文本到图像的生成依旧存在生成图像真实感缺失、无法保证给定文本和生成图像语义的一致性等问题。为了解决这两个问题,许多优秀学者运用不同模型结构取得了重大突破。经典的堆叠式模型包括StackGAN(stacked generative adversarial network)模型[5]、注意力生成对抗网络模型(fine-grained text to image generation with attentional generative adversarial networks, AttnGAN)[6]等;单阶段模型包括简单、有效的文本生成图像模型,深度融合生成对抗网络(deep-fusion generative adversarial networks, DF-GAN)[7]等。由于堆叠式模型不同生成器之间的纠缠,最终生成图像看起来像一个简单模糊的组合图。最重要的是,由于生成对抗网络的不稳定性,采用堆叠式模型会导致训练困难。

针对上述问题,文献[7]共同研究开发了一款简单、有效的文本生成图像模型DF-GAN。该模型是一种单阶段的文本生成图像模型,由一个生成器、一个判别器以及一个文本编码器组成,避免了不同生成器之间的纠缠,将GAN稳定训练到直接合成高分辨率图像。针对语义不一致的问题,DF-GAN的判别器由匹配感知梯度惩罚(matching-aware gradient penalty, MA-GP)和单向输出(one-way output)组成,主要作用是促使生成器合成更真实更符合语义一致性的图像。

为了进一步稳定DF-GAN模型训练,提高生成图片的质量,主要研究工作如下。

(1)增加条件增强模型,利用此模型提高训练过程的稳定性、增加生成图像的多样性。

(2)为了从文本分布出发拟合图像分布,增加视觉特征的多样性,扩大表现空间,在原有的DF-Block模块中增加一层Affine仿射块。

(3)为了提高文本与合成图像之间的语义一致性,在判别器中加入空间注意力模型[8]。

本研究提出的改善模型方法能更加准确地捕捉文本描述与生成图像的内在联系,在文本生成图像的初始得分(inception score, I_{IS})[9]和真实样本、生成样本在特征空间之间的距离(Fréchrt inception distance, F_{FID})[10]两项评价指标结果中取得了较大的提升。

1 相关研究

文本生成图像是指通过计算机程序生成自然语言描述的图像。这项技术的基本思想是将自然语言文本转化为计算机可处理的表示形式,并利用图像生成算法生成对应的图像。在深度学习技术出现之前,计算机视觉领域主要依赖传统方法生成图像,包括基于规则的方法、基于纹理的方法和基于统计的方法。基于规则的方法通常是手动设计的算法,通过对场景进行建模和分析生成图像,需要领域专家的知识和大量的人工劳动,因此限制了其应用范围。基于纹理的方法通过对现有的图像纹理进行分析,并在新的图像中重复使用这些纹理生成新的图像。这种方法通常能够产生逼真的图像,但存在难以处理复杂纹理和图案的局限性。基于统计的方法通过对大量图像进行统计分析学习图像的特征和分布,并利用这些信息生成新的图像。随着深度学习技术的发展,其应用范围和精度得到了大幅提升。文本生成图像技术的实现过程可以分为自然语言处理和图像生成两个主要步骤。自然语言处理的目标是将自然语言文本转换为计算机可处理的表示形式。目前常用的自然语言处理方法包括词向量表示、循环神经网络和变换器模型等。图像生成的目标是根据自然语言描述生成对应的图像。图像生成的方法包括基于卷积神经网络生成对抗网络、变分自编码器等。

文本描述生成相关图像跨模态任务最早是由文献[11]提出,采用GAN方式,由一个生成器和一个判别器构成,生成器负责生成相关的图像,判别器负责判断所生成图像的真假性,再反馈至生成器从而指导其工作,直到判别器判断不出生成图像的真假,达到一种拟合状态。利用生成对抗网络实现文本生成图像的任务也开始出现。2016年,文献[12]首次提出使用GAN-INT-CLS模型,分别在Oxford 102 Flowers[13]和Caltech-UCSD Birds-200-2011[14]数据集上使用简单的文本描述生成分辨率为64像素×64像素的图像,由于缺乏细节及生动性且图像局部纹理不清晰,GAN-INT-CLS模型比较容易出现过拟合现象,且该模型无法合成更高分辨率图像。为了能够生成更高分辨率及更高质量的图像,文献[5]提出了使用堆叠式的模型StackGAN分步实现并生成256像素×256像素的高分辨率、高质量图像。为了进一步提升生成图像的质量,文献[15]提出了StackGAN++。该模

型使用三级生成对抗网络,对生成器进行联合训练,并且以交替的方式对生成器和判别器进行训练。为了使生成图像更加细致、逼真,文献[6]提出 AttnGAN,可以通过关注文本描述中的相关词语,合成图像不同子区域的细粒度细节,此外,文献[6]还提出了一种深度注意多模态相似性模型,用于计算细粒度文本图像匹配损失,有利于生成更高质量图像。但是由于之前的生成对抗网络具有不可控性,如果改变句子中的某个单词,生成图像就会与原始文本生成的图像有较大不同。由此,文献[16]提出了既能有效生成高质量图像,又能根据自然语言描述控制部分图像生成的 ControlGAN 模型。每个单词在描述图像内容上都有不同的等级,而之前的方法仅使用同义的单词就能产生不同程度的影响,无法将单词与要生成的图像不同内容密切关联,且生成图像结果很大程度上取决于最初生成的图像质量。因此,文献[17]提出了动态记忆生成对抗网络生成高质量图像,该方法在初始生成图像质量不佳时,引入动态存储模块细化模糊图像内容。而现有能够生成高质量图像的生成对抗网络都采用堆叠式结构,致使生成的图像存在 3 个缺陷:(1)堆叠式的模型由于不同生成器之间的纠缠,最终细化的图像看起来像一个简单模糊的组合图。(2)现有研究通常会牺牲额外网络的文本对齐部分性能,以达到图像合成效果。(3)由于计算量大,跨模态注意力往往只能在分辨率为 64 像素×64 像素或 128 像素×128 像素的尺度上应用,限制了文本与图像融合的有效性。DF-GAN 只使用一个生成器、一个判别器和一个预训练过的文本编码器。DF-GAN 只引入了句子级的文本信息,限制了细粒度视觉特征合成能力并且缺乏视觉特征的多样性,并且生成对抗网络的训练普遍都具有不稳定性,而 DF-GAN 也具有这样的特性。此外,DF-GAN 模型忽略所生成图像与文本描述之间的语义一致性。

2　基于空间注意力及条件增强的文本生成图像网络模型

针对 DF-GAN 的不足,本研究做出相应的改进,以 DF-GAN 为基础,加入空间注意力机制和条件增强网络,改进后的模型称之为基于空间注意力及条件增强的文本生成图像模型(the text-to-image synthesis method based on spatial attention and conditional augmentation, ACDF-GAN),ACDF-GAN 模型结构如图 1 所示。

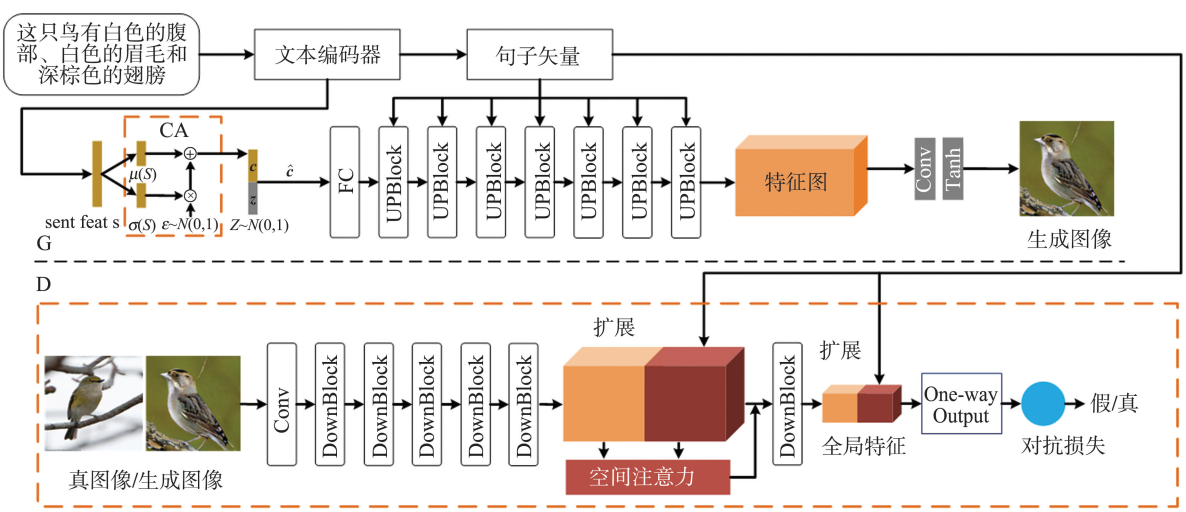


图 1　ACDF-GAN 模型结构

Fig.1　ACDF-GAN model structure

ACDF-GAN 由文本编码器、条件增强模块、生成器和一个具有空间注意力的匹配感知判别器组成。文本编码器是已经预训练好的且具有双向长短期记忆(long short-term memory, LSTM)[18]的编码器。本研究使用 AttnGAN 作为预训练模型。

ACDF-GAN 的生成器有两个输入:一是经过编码后的 Sentence vector;二是经过文本编码器后的文本描述通过条件增强模块,将文本描述转化为条件向量 $\boldsymbol{c}$,加入从正态分布中产生的随机噪声 $\boldsymbol{z}$。将条件向量 $\boldsymbol{c}$ 与随机噪声 $\boldsymbol{z}$ 拼接后的组合向量 $\hat{\boldsymbol{c}}$ 送入一个全连接层并重塑成所需要的尺寸,然后经过 7 个 UPBlock 层(上采样层、残差层[19]和 DF-Block 层)生成图像特征,最后经过卷积层将图像特征转化为图像。判别器经过 5 个 DownBlock 层把图像

转化为图像特征,将图像特征与文本特征扩张后经过空间注意力模块生成一个全局特征。运用 One Way Output 块计算对抗损失,最后判断生成的图像是否为真。

2.1 条件增强模型

堆叠式模型 StackGAN 是条件增强技术的首次应用。条件增强技术是通过文本编码器后的句子特征 $\boldsymbol{s}$,由 $\boldsymbol{s}$ 的高斯分布 $N(\boldsymbol{\mu}(\boldsymbol{s}),\sum(\boldsymbol{s}))$ 得到平均协方差矩阵 $\boldsymbol{\mu}(\boldsymbol{s})$ 和对角协方差矩阵 $\boldsymbol{\sigma}(\boldsymbol{s})$,条件向量

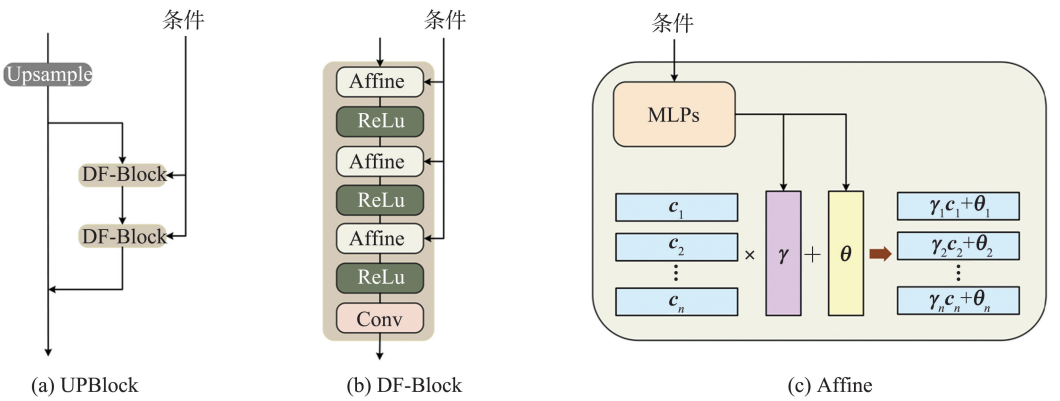
$$\boldsymbol{c}=\boldsymbol{\mu}(\boldsymbol{s})+\boldsymbol{\sigma}(\boldsymbol{s})\boldsymbol{\varepsilon},$$

式中,$\boldsymbol{\varepsilon}\sim\boldsymbol{N}(0,\boldsymbol{I})$,其中 $\boldsymbol{N}(0,\boldsymbol{I})$ 为均值为 0、协方差矩阵为单位矩阵 $\boldsymbol{I}$ 的正态分布(或高斯分布)。

通过组合向量的方式不仅能在一定程度上缓解图像文本匹配时数据的压力,且有助于增强对条件流形上小扰动的鲁棒性。

2.2 生成器模型

生成器由 7 个 UPBlocks 组成。DF-Block 是原有模型 DF-GAN 的一种深度文本图像融合块,DF-Block 叠加了多个 Affine 仿射块和 ReLU 层。UPBlock 结构及内部图如图 2 所示。



(a) UPBlock (b) DF-Block (c) Affine

图 2 UPBlock 结构及内部图

Fig.2 UPBlock structure and internal diagram

由图 2(a)(b)可知,UPBlock 有两块 DF-Block。DF-Block 由 3 个 Affine 仿射块、3 个 ReLu 激活层和一个卷积层构成。这种方法受到了条件批量规范化以及自适应实例规范化的启发,在两个 Affine 仿射块中添加一个 ReLu 激活层,将非线性引入融合过程,使文本和图像特征融合时能够更加充分利用文本信息,扩大了融合模块的表示空间,有利于从不同的文本描述中生成语义一致的图像。图2(c)是 Affine 仿射块,由两个多层感知器组成,多层感知器预测语言条件下的通道尺度参数 $\boldsymbol{\gamma}$ 和预其移位参数 $\boldsymbol{\theta}$,其表达式为:

$$\boldsymbol{\gamma}=M_{\mathrm{LP1}}(\boldsymbol{e}),$$

$$\boldsymbol{\theta}=M_{\mathrm{LP2}}(\boldsymbol{e}),$$

式中,$\boldsymbol{e}$ 为句子向量,M_{LP} 为多层感知器。

Affine 仿射块使用 $\boldsymbol{\gamma}$ 进行通道方向的标度运算,使用 $\boldsymbol{\theta}$ 进行通道方向的移位运算,表达式为:

$$\mathrm{Affine}(\boldsymbol{x}_i\mid\boldsymbol{e})=\boldsymbol{\gamma}_i\boldsymbol{x}_i+\boldsymbol{\theta}_i,$$

式中,$\boldsymbol{x}_i$ 为视觉特征图的第 i 个通道信息,$\boldsymbol{\gamma}_i$ 和 $\boldsymbol{\theta}_i$ 分别是为视觉特征图第 i 通道的缩放参数和位移参数。

2.3 具有空间注意力的匹配感知判别器

判别器由匹配感知梯度惩罚、单向输出和空间注意力组成。在判别器中加入空间注意力,可以促进生成器生成更真实、更符合文本语义一致性的图像。如图 1 所示,5 个 DownBlock 用于将图像转化为图像特征。结合图像特征映射 $\boldsymbol{P}$ 与句子特征 $\boldsymbol{s}$ 中的信息,空间注意力生成一个注意力映射 $\boldsymbol{\alpha}$,$\boldsymbol{\alpha}$ 可以抑制无关区域的句子向量,表达式为:

$$\boldsymbol{x}_{w,h}=M_{\mathrm{LP}}(\boldsymbol{P}_{w,h},\boldsymbol{s}),$$

$$\boldsymbol{\alpha}_{w,h}=\frac{1}{1+\boldsymbol{e}^{-x_{w,h}}}\Big/\sum_{w=1,h=1}^{w,h}\frac{1}{1+\boldsymbol{e}^{-x_{w,h}}},$$

$$\boldsymbol{S}_{w,h}=\boldsymbol{s}\times\boldsymbol{\alpha}_{w,h},$$

式中:$\boldsymbol{P}_{w,h}$ 为图像的图像特征,在坐标轴的表示方式为 (w,h);$\boldsymbol{\alpha}_{w,h}$ 为注意力概率,是将计算出的 $\boldsymbol{x}_{w,h}$ 通过计算权重转换成注意概率;$\boldsymbol{S}_{w,h}$ 为句子特征匹配图像特征的权重。

为了稳定 GAN 训练,在计算 $\boldsymbol{\alpha}_{w,h}$ 时使用软阈值函数方法:

$$\boldsymbol{P}(\boldsymbol{x}_k)=\frac{1}{1+\boldsymbol{e}^{-x_k}}\Big/\sum_{j-1}^{k}\frac{1}{1+\boldsymbol{e}^{-x_j}},\tag{1}$$

式(1)未采用当前流行的 softmax 函数,可以使最大概率最大化,并且抑制其他概率接近 0。极小的概率阻碍了梯度的反向传播,从而加剧 GAN 训练的不稳定性。通过软阈值函数可以防止注意概率

接近 0,并且可以提高反向传播的效率。空间注意力模型将更多的文本特征分配给相关的图像区域,这有助于判别器确定文本、图像是否相匹配。

2.4　损失函数

判别器的训练目标将生成器生成的不匹配图像作为负样本,进而监督生成器生成关联性更强的图像。在给定文本和匹配文本上使用 hinge loss[20] 的匹配感知梯度惩罚(mastching-aware gradient penalty, MA-GP)作为损失函数。其判别器的损失函数为:

$$L_{\text{adv}}^{\text{D}}=E_{\boldsymbol{x}\sim\boldsymbol{p}_{\text{data}}}[\max(0,1-\boldsymbol{D}(\boldsymbol{x},\boldsymbol{t}))]+\frac{1}{2}E_{\boldsymbol{x}\sim\boldsymbol{p}_{\text{G}}}[\max(0,1-\boldsymbol{D}(\hat{\boldsymbol{x}},\boldsymbol{t}))]+\frac{1}{2}E_{\boldsymbol{x}\sim\boldsymbol{p}_{\text{data}}}[\max(0,1-\boldsymbol{D}(\boldsymbol{x},\hat{\boldsymbol{t}}))],$$

式中,$\boldsymbol{t}$ 为给定的文本描述,$\hat{\boldsymbol{t}}$ 为不匹配的文本描述,$\boldsymbol{x}$ 为真实的图像,$\hat{\boldsymbol{x}}$ 为生成的图像,$E_{x\sim p_{\text{data}}}$ 为真实图像样本平均值,$E_{x\sim p_{\text{G}}}$ 为生成图像样本平均值。

生成器的损失函数为:

$$L_{\text{adv}}^{\text{G}}=E_{\boldsymbol{x}\sim\boldsymbol{p}_{\text{data}}}[\min(\boldsymbol{D}(\boldsymbol{x},\boldsymbol{t}))]。$$

3　试验结果与分析

3.1　试验环境及数据集

本研究采用硬件环境如下。

CPU:Intel(R) Core(M) i5-12400f、六核心、十二线程。最大睿频频率 4.4 GHz。

GPU:NVIDIA GeForce RTX3090,其 CUDA 核心数量为 10 496 个,显存为 24 G。操作系统为 64 位的 Ubuntu18.04、CUDA Toolkit11.3、Python 3.8,深度学习框架为 Pytorch1.9,该框架在 GPU 上运行。

本次试验所使用的数据集为公开数据集 Caltech-UCSD Birds-200-2011(CUB)、Oxford 102 Flowers 以及 COCO[21] 数据集。

3.2　试验设置

本次试验训练阶段网络优化器采用文献[22]提出的 Adam。根据双时间刻度更新规则(two timescale update rule, TTUR),生成器的学习率设置为 0.000 1,判别器的学习率设置为 0.000 4,训练中的 Batch Size 设置为 16。为了更好地与 DF-GAN 做比较,需与 DF-GAN 保持相同的训练轮数。根据数据集的不同,训练轮数也不同,CUB 数据集的训练轮数设置为 600 轮,Oxford 102 Flowers 数据集的训练轮数设置为 600 轮,COCO 数据集训练轮数设置为 120 轮。

优化器 Adam 的 β_1 设置为 0.0,β_2 设置为 0.9。β_1 和 β_2 为两个指数加权平均值的衰减系数。

3.3　评价指标

为了量化本研究的试验结果,选择 I_{IS} 和 F_{FID} 作为 ACDF-GAN 模型性能的评价指标。

I_{IS} 是评价图像生成领域的一个重要指标,将清晰度和多样性作为图像生成效果的评价指标,表示为:

$$I_{\text{IS}}=\exp(E_{\boldsymbol{x}}\boldsymbol{K}_{\text{L}}(\boldsymbol{p}(\boldsymbol{y}|\boldsymbol{k})\parallel\boldsymbol{p}(\boldsymbol{y}))),$$

式中:$\boldsymbol{k}$ 为一个生成的样本,表示模型生成的图像;$\boldsymbol{y}$ 为预测的标签,表示这个图像包含的主要物体;$\boldsymbol{p}(\boldsymbol{y}|\boldsymbol{x})$ 为条件概率,代表给出一个图像,预测图像中包含的物体的概率,即有把握对图像进行正确分类;$\boldsymbol{p}(\boldsymbol{y})$ 为边缘概率,即标签的分布情况;$E_{\boldsymbol{x}}$ 表示遍历所有的生成样本,求平均值;$\boldsymbol{K}_{\text{L}}$ 表示散度。

在一定的程度上,$\boldsymbol{p}(\boldsymbol{y}|\boldsymbol{x})$ 代表图像的质量,概率越高越好。这里我们希望标签分布均匀,而不希望模型生成的都是某一类的图像。因此,$\boldsymbol{p}(\boldsymbol{y})$ 可以代表模型生成图像的多样性。

$\boldsymbol{K}_{\text{L}}$ 散度与这两者之间的关系为:条件概率 $\boldsymbol{p}(\boldsymbol{y}|\boldsymbol{x})$ 越小,边缘概率 $\boldsymbol{p}(\boldsymbol{y})$ 越高,则 $\boldsymbol{K}_{\text{L}}$ 散度相对越大,所生成图像的多样性会增加,并且质量也会越高。由此可知,I_{IS} 越大,生成图像的效果越好,图像就会越清晰,质量就会越高,就更具多样性。

F_{FID} 根据预训练网络提取特征,测量真实图像分布与生成图像分布之间的距离,其表达式为:

$$F_{\text{FID}}=\parallel\boldsymbol{\mu}_{\text{r}}-\boldsymbol{\mu}_{\text{g}}\parallel_2^2+\boldsymbol{T}_{\text{r}}\left(\sum_{\text{r}}+\sum_{\text{g}}-2\left(\sum_{\text{r}}\sum_{\text{g}}\right)^{\frac{1}{2}}\right),$$

式中,$\boldsymbol{\mu}$ 为分布的均值,Σ 为协方差,$\boldsymbol{T}_{\text{r}}$ 为迹(矩阵对角线上的元素和),$\boldsymbol{\mu}_{\text{r}}$ 为真实样本特征均值,$\boldsymbol{\mu}_{\text{g}}$ 为生成样本特征均值,$\sum_{\text{r}}$ 和 $\sum_{\text{g}}$ 为真实与生成图像特征的协方差。

在最佳情况下,F_{FID} 得分为 0.0,表示两组图像相同,因此 F_{FID} 越小越好。

3.4　定量与定性结果分析

3.4.1　定量结果分析

本研究在 CUB、Oxford 102 Flowers 以及 COCO 数据集上进行了测试,随机产生了大约 30 000 张图像,分别计算了 I_{IS} 和 F_{FID},并且与当前主流的 StackGAN++[15]、AttnGAN[6]、MirrorGAN[23]、DM-GAN[17] 和 DF-GAN 模型进行了比较,结果如表 1 所示。

表 1 不同模型方法在 CUB、COCO、Oxford 102 Flowers 数据集上 I_{IS}、F_{FID}

Table 1 I_{IS}, F_{FID} on CUB, COCO, Oxford 102 Flowers dataset for different model approaches

模型方法	I_{IS}		F_{FID}	
	CUB	Oxford	CUB	COCO
StackGAN++	4.04±0.06	3.26	15.30	81.59
AttnGAN	4.36±0.03	—	15.38	35.49
MirrorGAN	4.56±0.05	—	18.34	34.71
DM-GAN	4.75±0.07	—	16.09	32.64
DF-GAN	4.86±0.04	3.80	14.81	21.63
本研究	**4.96±0.10**	**3.90**	**11.74**	**19.63**

由表 1 可知，与之前的 DF-GAN 相比，在 CUB 数据集上，本研究方法 I_{IS} 提高了 2.05%，在 Oxford 102 Flowers 数据集上提高了 2.63%。在 CUB 数据集与 COCO 数据集上，F_{FID} 分别降低了 20.73%、9.25%。本研究方法的 I_{IS} 和 F_{FID} 均优于目前其他主流网络。

3.4.2 定性结果分析

为了更直观对比不同模型以及数据集上结果，本节从视觉结果角度观察生成图像的质量。4 种不同 GAN 模型在 CUB 数据集及 COCO 数据集的视觉结果如图 3 所示。

(a) CUB数据集 (b) COCO数据集

图 3 4 种不同的 GAN 模型在 CUB 数据集及 COCO 数据集的视觉结果

Fig.3 Visual results of 4 different GAN models on CUB dataset and COCO dataset

由图 3(a)可知，与其他模型对比，本研究模型在 CUB 数据集上纹理特征和细节方面的表现更好，比如羽毛、眼睛和爪子既丰富又生动，并且与文本描述相符合。本研究模型具有空间注意力，所以生成图像的相关内容更多、扭曲部分更少。由图 3(b)可知，在更为复杂的 COCO 数据集上，图像依然具有清晰的纹理和丰富的色彩。由此可以证明，本研究生成图像相较其他 3 种模型在语义一致性和图像质量方面都有极大提升。

3.5 消融试验

为了能够验证条件增强模块与空间注意力模块的有效性，分别设置 DF-GAN、DF-GAN-CA、DF-GAN-ST 和 DF-GAN-CA-ST 4 组对比试验，其中 DF-GAN 为基础网络，CA（conditional augmentation）为条件增强模块，ST 为空间注意力模块。在 CUB 数据集试验结果如表 2 所示。

表 2 消融试验结果对比

Table 2 Comparison of results of ablation experiments

模块	I_{IS}	F_{FID}
DF-GAN	4.86±0.04	14.81
DF-GAN-CA	4.91±0.08	12.60
DF-GAN-ST	4.69±0.05	11.57
DF-GAN-CA-ST	**4.96±0.10**	**11.74**

由表 2 可知，条件增强模块与空间注意力模块对图像的生成结果均有正向调节作用，将两个模块相结合可达到本试验的最佳效果。由此证明了本研究方法的有效性。空间注意力的可视化结果如图 4 所示。

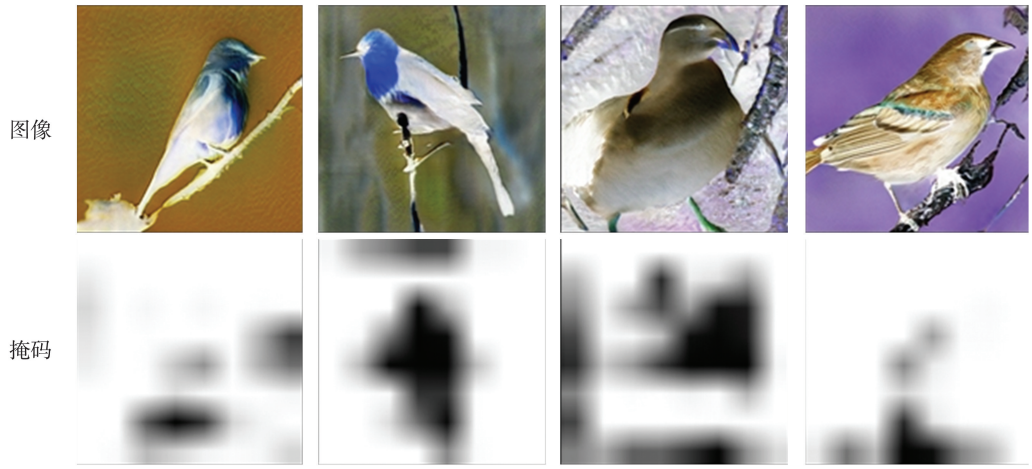


图 4　空间注意力的可视化结果
Fig.4　Visualization of spatial attention results

由图 4 可以直观看出空间注意力能识别与标题相关的区域,从而使判别器能够在图像与标题之间做出更好的比较。

4　结束语

本研究对一种简单有效的文本生成图像基准模型 DF-GAN 进行了改进。通过引入条件增强模型,提高了训练过程的稳定性,增加了生成图像的多样性。为了扩大表现空间,在原有的 DF-Block 模块中增加一层 Affine 仿射块。为了提高语义一致性,在判别器中加入了空间注意力模型。试验结果表明,本研究模型生成的图像在 I_{IS} 和 F_{FID} 都取得了较好效果。但是,由于本研究网络模型较大,导致训练时间长,一些文本类别生成的图像扭曲,在一些语义的细节上存在偏差,需要在之后的工作中继续优化模型,完善模型结构。

参考文献:

[1] YI X, WALIA E, BABYN P. Generative adversarial network in medical imaging: a review[J]. Medical Image Analysis, 2019, 58:101552.

[2] 胡名起. 基于生成对抗网络的文本生成图像研究[D]. 南京:东南大学, 2020.
HU Mingqi. Research on text-to-image generation based on generative adversarial network[D]. Nanjing: Southeast University, 2020.

[3] GOLDBERG Y. Neural network methods for natural language processing[M]. Berlin: Springer Nature, 2022.

[4] XU K, BA J, KIROS R, et al. Show, attend and tell: neural image caption generation with visual attention[C]//Proceedings of International Conference on Machine Learning. Lille, France: PMLR, 2015: 2048-2057.

[5] ZHANG H, XU T, LI H S, et al. StackGAN: realistic image synthesis with stacked generative adversarial networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(8): 1947-1962

[6] XU T, ZHANG P, HUANG Q, et al. AttnGAN: fine-grained text to image generation with attentional generative adversarial networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Dhaka, Bangladesh: IEEE Press, 2018: 1316-1324.

[7] TAO M, TANG H, WU S, et al. DF-GAN: deep fusion generative adversarial networks for text-to-image synthesis[EB/OL]. (2020-08-13)[2023-03-18]. https://arxiv.org/abs/2008.05865v1.

[8] DU C, ZHANG L, SUN X, et al. Enhanced multi-channel feature synthesis for hand gesture recognition based on CNN with a channel and spatial attention mechanism[J]. IEEE Access, 2020, 8: 144610-144620.

[9] SALIMANS T, GOODFELLOW I, ZAREMBA W, et al. Improved techniques for training GANs[J]. Advances in Neural Information Processing Systems, 2016, 29(2): 2234-2242.

[10] HEUSEL M, RAMSAUER H, UNTERTHINER T, et al. GANs trained by a two time-scale update rule converge to a local Nash equilibrium[J]. Advances in Neural Information Processing Systems, 2017, 30(4): 6629-6640.

[11] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.

[12] REED S, AKATA Z, YAN X, et al. Generative adversarial text to image synthesis[C]// Proceedings of International Conference on Machine Learning. Lille, France: PMLR, 2016: 1060-1069.

[13] NILSBACK M E, ZISSERMAN A. Automated flower

classification over a large number of classes[C]//Proceedings of 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing. Bhubaneswar, India: IEEE Press, 2008:722-729.

[14] WAH C, BRANSON S, WELINDER P, et al. The Caltech-UCSD birds-200-2011 dataset[J]. California Institute of Technology, 2011, 7(1): 1-8.

[15] ZHANG H, XU T, LI H S, et al. StackGAN: text to photo-realistic image synthesis with stacked generative adversarial networks[C]//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE Press, 2017: 5908-5916.

[16] LI B, QI X, LUKASIEWICZ T, et al. Controllable text-to-image generation[J]. Advances in Neural Information Processing Systems, 2019, 32(3):2065-2075.

[17] ZHU M F, PAN P B, CHEN W, et al. DM-GAN: dynamic memory generative adversarial networks for text-to-image synthesis[C]// Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE Press, 2019: 5795-5803.

[18] SCHUSTER M, PALIWAL K K. Bidirectional recurrent neural networks[J]. IEEE transactions on Signal Processing, 1997, 45(11): 2673-2681.

[19] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE Press, 2016: 770-778.

[20] XUE W, ZHONG P, ZHANG W, et al. Sample-based online learning for bi-regular hinge loss[J]. International Journal of Machine Learning and Cybernetics, 2021, 12: 1753-1768.

[21] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context[C]// Proceedings of Computer Vision-ECCV 2014. Zurich, Switzerland: Springer, 2014: 740-755.

[22] KINGMA D P, BA J. Adam: a method for stochastic optimization[EB/OL]. (2014-12-22)[2023-03-18]. https://arxiv.org/abs/1412.6980.

[23] QIAO T, ZHANG J, XU D, et al. MirrorGAN: learning text-to-image generation by redescription[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Changsha, China: IEEE Press, 2019: 1505-1514.

(编辑:李骏)

(上接第 48 页)

[23] 杨习贝, 颜旭, 徐苏平, 等. 基于样本选择的启发式属性约简方法研究[J]. 计算机科学, 2016, 43(1): 40-43.
YANG Xibei, YAN Xu, XU Suping, et al. New heuristic attribute reduction algorithm based on sample selection[J]. Computer Science, 2016, 43(1): 40-43.

[24] XIAO Z, CHANG L M, DE G C, et al. Feature selection in mixed data: a method using a novel fuzzy rough set-based information entropy[J]. Pattern Recognition, 2016, 56(1): 1-15.

[25] WANG C Z, HU Q H, WANG X Z, et al. Feature selection based on neighborhood discrimination index[J]. IEEE Trans on Neural Networks and Learning Systems, 2018, 29(7): 2986-2999.

[26] 张昭琴, 徐泰华, 鞠恒荣, 等. 基于粒度的加速求解约简策略[J]. 南京理工大学学报, 2021, 45(4): 401-408.
ZHANG Zhaoqin, XU Taihua, JU Hengrong, et al. Accelerative solving reduct strategy based on granularity[J]. Journal of Nanjing University of Science and Technology, 2021, 45(4): 401-408.

(编辑:孙亚彤)