

基于特征融合的文本到图像生成方法

周 杭¹, 方清茂², 张 美², 卿粼波¹, 何小海¹

(1. 四川大学电子信息学院, 四川成都 610065; 2. 四川省中医药科学院, 四川成都 610041)

摘要:近年来,随着生成对抗网络(Generative Adversarial Network, GAN)技术的不断发展,其被广泛应用于文本生成图像的任务中。现单阶段生成对抗模型大多只使用了句子文本描述,没有充分利用文本信息。为此,本文以单阶段生成对抗模型为基础,提出了一种基于特征融合的文本到图像生成方法(FFGAN)。一方面,构建文本-图像跨模态融合模块使单词向量特征和图像特征能够有效融合,丰富生成图像的细节;另一方面,引入感知损失来缩小生成图像和目标图像的距离,使得图像逼真度更高。实验结果表明,在CUB数据集上,FFGAN模型的IS分数达到了 5.22 ± 0.08 ,FID分数达到了13.91。在COCO数据集上,FFGAN模型的FID分数达到了16.97。大量实验充分证明了FFGAN的优越性以及有效性。

关键词:文本生成图像;生成对抗网络;跨模态融合;感知损失

本文著录格式:周杭,方清茂,张美,等.基于特征融合的文本到图像生成方法[J].新一代信息技术,2023,6(15):06-12

中图分类号:TP391.41

文献标识码:A

Text to Image Generation Method Based on Feature Fusion

ZHOU Hang¹, FANG Qing-mao², ZHANG Mei², QING Lin-bo¹, HE Xiao-hai¹

(1. School of Electronic Information, Sichuan University, Chengdu, Sichuan 610065, China; 2. Sichuan Academy of Traditional Chinese Medicine, Chengdu, Sichuan 610041, China)

Abstract: In recent years, with the continuous development of generative adversarial network (GAN) technology, it has been widely used in the task of generating images from text. Most existing single-stage GAN solely rely on textual sentences, failing to fully leverage the available textual information. To address these limitations, this study proposes a feature fusion-based text-to-image generation method (FFGAN), based on a single-stage GAN. FFGAN incorporates a text-image cross-modal fusion module, enabling effective fusion of word vector features and image features. This fusion enriches the generated image's details. Additionally, perceptual loss is introduced to minimize the discrepancy between the generated and target images, enhancing the realism of the generated image. Experimental results on the CUB dataset demonstrate that the FFGAN model achieves an IS score of 5.22 ± 0.08 and an FID score of 13.91. On the COCO dataset, the FID score of the FFGAN model reaches 16.97. Through numerous experiments, FFGAN's superiority and effectiveness have been conclusively demonstrated.

Key words: text to image generation; generative adversarial networks; cross-modal fusion; perceptual loss

Citation: ZHOU Hang, FANG Qing-mao, ZHANG Mei, et al. Text to Image Generation Method Based on Feature Fusion[J]. New Generation of Information Technology, 2023, 6(15): 06-12

1 引言

近年来,随着Goodfellow等^[1]首次提出生成对抗网络(Generative Adversarial Network, GAN)之后,文本生成图像、图像修复^[2]、图像风格迁移^[3]等任务得到了快速的发展。文本生成图像任务旨在利用文本描述信息生成与文本信息相对应的图像,是当前众多研究

者研究的热门课题之一。2016年,Reed等^[4]提出了GAN-INT-CLS模型,该模型中首次添加文本信息作为生成网络和判别网络的约束条件。后来,Zhang等^[5]提出了StackGAN模型,该模型主要由两个阶段构成,其中第一个阶段利用文本描述生成 64×64 大小的低分辨率图像,然后第二个阶段利用前面生成的低分辨率

基金项目:成都市重大科技应用示范项目(No.2019-YF09-00120-SN)

通讯作者:何小海

图像通过上采样得到 256×256 大小的图像。2018 年, Xu 等^[6]提出了 AttnGAN 模型,该模型包含了一种注意生成网络,其能够通过文本描述中的单词文本在图像的子区域中合成更加细粒度的细节,而且该模型还提出了深度注意多模态相似模型用于计算文本和图像之间的匹配损失,从而指导生成器的训练。Zhu 等^[7]提出的 DM-GAN 模型中引入了一个记忆写入门来动态地选择与生成图像相关的单词,然后利用图像中所包含的信息来确定每个单词的重要性,从而使生成图像更好地与文本描述相对应。由于前面的堆叠式 GAN 网络都是先利用文本描述生成低分辨率的图像,之后再不断地对低分辨率图像进行细化从而得到高分辨率图像,所以最终生成的高分辨率图像质量取决于第一阶段生成的初始图像。即如果第一阶段生成的初始图像效果不好,那么最终所生成的图像效果也会较差。并且由于堆叠式 GAN 引入了不同生成器之间的纠缠,使得最终的图像看起来像是模糊图像和一些细节的简单组合。针对上述问题,DFGAN^[8]以及 SSAGAN^[9]等提出了基于单个生成器和鉴别器的单阶段生成网络。Tao 等^[8]提出的 DF-GAN 模型是一个单阶段生成对抗模型,该模型中包含了一个深度文本图像融合模块:通过一系列仿射变换实现文本图像的融合,并且还设计了一个由匹配感知梯度惩罚和单向输出组成的目标感知鉴别器来增强文本-图像语义的一致性。Liao 等^[9]提出的 SSAGAN 模型中设计了一种语义空间感知模块来深入地融合文本和图像特征。近来,由谷歌发布的 Imagen^[10]以及 OpenAI 发布的 DALL-E-2^[11]等扩散模型在文本生成图像方向上取得了卓越的性能。扩散模型基于马尔科夫链建模,其基本原理是在前向过程中对图像逐步添加噪声,直到图像被完全破坏成高斯噪声,然后在逆向过程学习从高斯噪声恢复为原始图像。虽然扩散模型展现出了很强的性能,但是由于其耗时的推理过程以及巨大的训练成本,所以基于 GAN 的文本生成图像任务仍然具有广阔的研究前景。

虽然 DFGAN 模型和 SSAGAN 模型在文本生成图像任务中取得了不错的效果,但是它们在文本特征和图像特征融合的过程中只使用了句子向量特征,并没有使用比句子向量特征更加细粒度的单词向量特征,所以使得生成图像的细节并不是很完善^[12]。为了解决 SSAGAN 此类单阶段生成网络存在的问题,本文提出了一种基于特征融合的文本到图像生成方法——FFGAN。该方法以 SSAGAN 作为基线模型,针对生成图像细节不够完善的问题,一方面通过引入单词文本特征,构建文本-图像跨模态融合模块实现单词向量特

征和图像特征的跨模态融合,即通过单词向量特征去指导图像特征的注意力学习,从而达到利用单词向量特征不断丰富图像特征的目的。另一方面,引入感知损失来缩小生成图像特征与原始图像特征之间的距离,从而使得生成图像更加真实。本文在 CUB 数据集和 COCO 数据集上进行了实验,实验结果表明,FFGAN 模型具有一定的优越性。

2 FFGAN 模型

2.1 模型整体结构

本文提出的 FFGAN 模型的整体结构图如图 1 所示。首先通过 Bi-LSTM 文本编码器对输入的文本描述进行特征编码,分别生成句子向量特征和单词向量特征,通过语义空间感知模块结合句子向量特征和随机噪声生成粗粒度的图像特征,然后利用文本图像跨模态融合模块对生成的粗粒度图像特征以及单词向量特征进行跨模态融合,从而生成更加细粒度的图像特征,最后提出了包含对抗性损失、文本图像匹配损失以及感知损失的生成器损失模块,能够有效提升图像的真实性。

2.2 文本编码器

在文本生成图像任务中,文本编码器起着至关重要的作用。本文采用的文本编码器是双向长短期记忆卷积神经网络(Bi-directional Long Short-Term Memory, Bi-LSTM)。首先将文本描述送入到文本编码器,文本编码器会将文本描述编码成全局句子特征向量和单词特征向量,分别表示为 $s \in \mathbb{R}^D$, $w \in \mathbb{R}^{D \times N}$ 。其中 D 表示句子特征向量以及单词特征向量的维度, N 表示单词的个数。

2.3 语义空间感知模块

语义空间感知模块由上采样模块、语义掩码预测器以及具有残差连接的语义空间条件批量归一化模块构成。其结构图如图 2 所示。

上采样模块:采用双线性插值方法对图像特征图进行扩张。

掩码预测器:对上采样之后的图像特征进行一系列的卷积、批量归一化等操作之后,来预测当前图像特征图的语义掩码图。其能够直观地展示当前图像特征图哪些部分还需要通过文本信息去增强,从而实现细化后的图像特征图在语义上和给定的文本描述更加地一致。

语义空间条件批量归一化模块:将掩码预测器输出的掩码特征图作为语义条件批量归一化的空间信

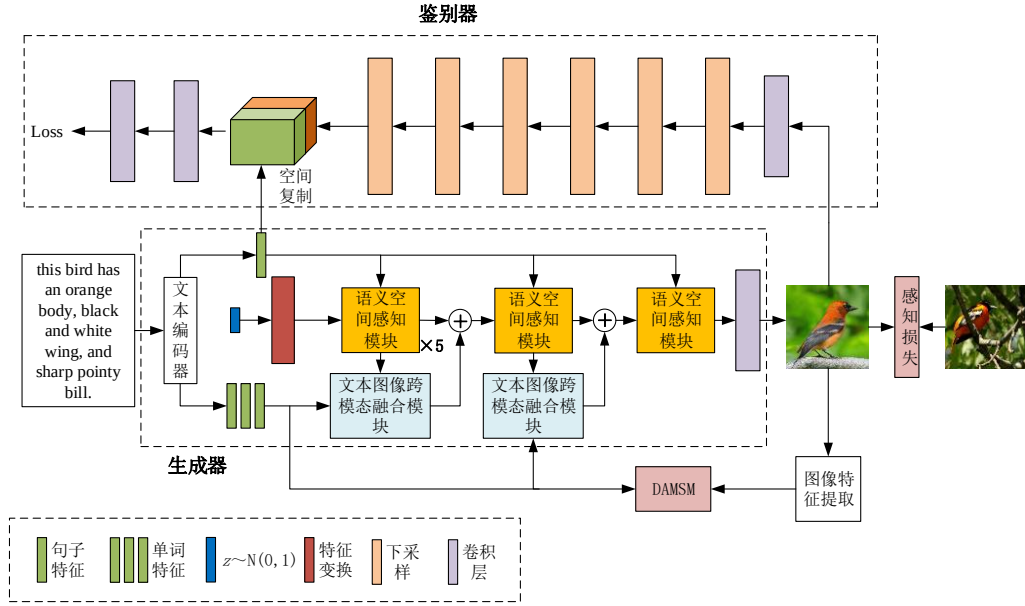


图1 FFGAN 模型框架图

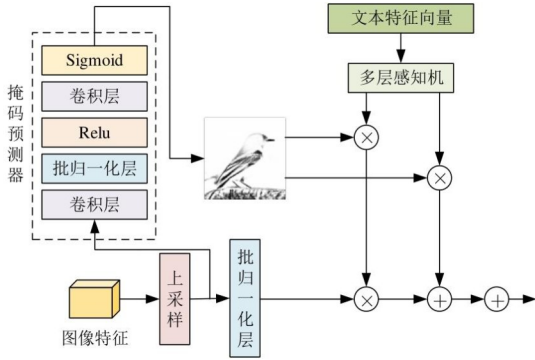


图2 语义空间感知模块

息,从而确定在图像特征图上添加文本信息的位置以及添加文本信息的量。如式(1)所示:

$$\tilde{\mathbf{x}}_{\text{ncw}} = \mathbf{m}_{i(h,w)}(\gamma_c(\bar{\mathbf{e}})\hat{\mathbf{x}}_{\text{ncw}} + \beta_c(\bar{\mathbf{e}})) \quad (1)$$

其中, $\mathbf{m}_{i(h,w)}$ 表示掩码预测器输出的掩码特征图特征, $\bar{\mathbf{e}}$ 表示句子向量特征, $\gamma_c(\bar{\mathbf{e}})$ 以及 $\beta_c(\bar{\mathbf{e}})$ 表示从文本向量特征中学习到的调制参数, $\hat{\mathbf{x}}_{\text{ncw}}$ 表示输入的图像特征, $\tilde{\mathbf{x}}_{\text{ncw}}$ 表示经过语义空间条件批量归一化之后的图像特征。

最后将得到的句子文本信息增强之后的图像特征与输入的图像特征进行残差相加,从而得到新的图像特征,如式(2)所示:

$$\mathbf{x}_{\text{ncw}} = \hat{\mathbf{x}}_{\text{ncw}} + \lambda_1 \tilde{\mathbf{x}}_{\text{ncw}} \quad (2)$$

其中, λ_1 表示网络自动学习并优化的权重系数, $\mathbf{x}_{\text{ncw}}$ 表示经过残差相加之后的特征。

2.4 文本-图像跨模态融合模块

文本-图像跨模态融合模块由单词-图像对齐模块以及单词-图像融合模块组成,文本-图像跨模态融合

模块结构图如图3所示。

由于单词特征向量和图像特征是两种不同模态以及不同维度的特征,所以想要实现文本-图像的跨模态融合,那么就需要解决文本和图像之间数据异构的问题。针对该问题,引入单词-图像对齐模块。输入的单词特征向量为 $\mathbf{w} \in \mathbb{R}^{D \times N}$, 输入的图像特征为 $\mathbf{h} \in \mathbb{R}^{C \times H \times W}$, 其中 D 表示单词特征向量的维度, N 表示单词个数, C 表示图像特征的通道数, $H \times W$ 表示图像特征图的大小。将 $\mathbf{w}$ 和 $\mathbf{w}^T$ 相乘,得到新的固定大小的单词特征 $\mathbf{w}_1$, 其公式可以用式(3)表示:

$$\mathbf{w}_1 = \mathbf{w} \times \mathbf{w}^T \quad (3)$$

其中, $\mathbf{w}^T \in \mathbb{R}^{N \times D}$, $\mathbf{w}_1 \in \mathbb{R}^{D \times D}$ 。

然后对单词特征向量进行尺度变换大小,得到单词特征 $\mathbf{w}_2$ 。其公式可以用式(4)表示:

$$\mathbf{w}_2 = \text{Reshape}(\mathbf{w}_1) \quad (4)$$

其中, $\mathbf{w}_2 \in \mathbb{R}^{C \times S_0}$, S_0 表示 $D \times D \div C$ 。

再通过线性层将单词特征 $\mathbf{w}_2$ 转换为新的单词特征 $\mathbf{w}_3$, 其公式可以用式(5)表示:

$$\mathbf{w}_3 = \text{Linear}(\mathbf{w}_2) \quad (5)$$

其中, $\mathbf{w}_3 \in \mathbb{R}^{C \times S_1}$, $S_1 = H \times W$ 。

当单词特征向量和图像特征转换到同一个维度之后,对图像特征和单词特征向量进行跨模态融合。单词-图像融合模块旨在利用单词特征向量去指导图像特征的注意力学习,从而达到细化图像的目的。首先输入单词特征向量大小为 $\mathbf{w}_3$, 图像特征大小为 $\mathbf{h}_1 \in \mathbb{R}^{C \times S_1}$, 将单词特征通过全连接层映射到 $\mathbf{k}$ 、 $\mathbf{v}$ 特征向量空间,将图像特征通过全连接层映射到 $\mathbf{q}$ 特征向量空间。如式(6)、式(7)所示:

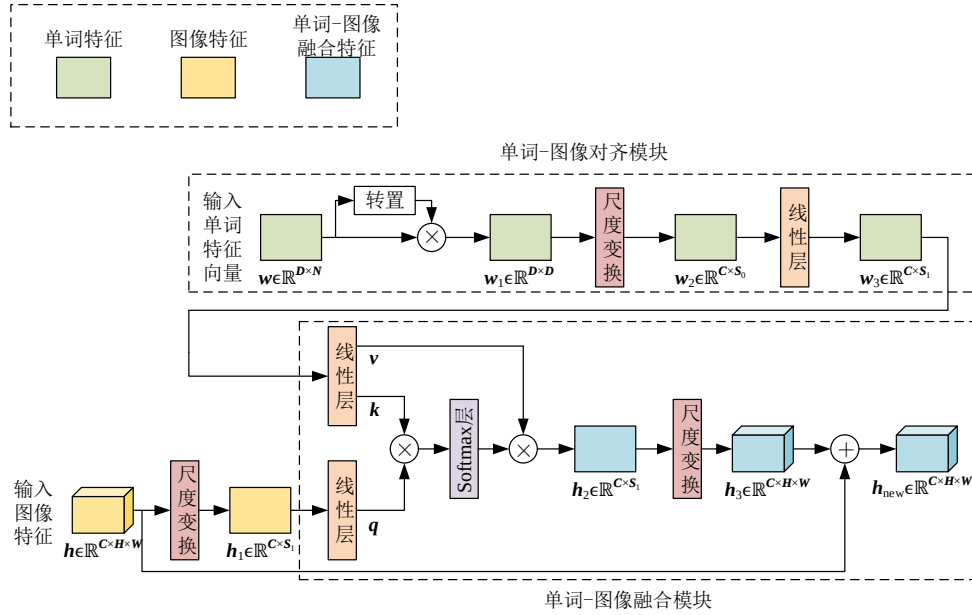


图3 文本-图像跨模态融合模块

$$q = \text{Linear}(h_1) \quad (6)$$

$$k, v = \text{Linear}(w_3) \quad (7)$$

通过 q 特征向量以及 k 特征向量计算相似度概率,最后与 v 进行相乘,得到最终的注意力分数。实现了单词-图像特征的跨模态融合之后,如式(8)所示:

$$\text{Attention}(q, k, v) = \text{Softmax}\left(\frac{qk^T}{\sqrt{d_k}}\right)v \quad (8)$$

最后将融合之后的特征进行结构变换,得到新的融合之后的图像特征 $h_3 \in \mathbb{R}^{C \times H \times W}$,并和之前的图像特征进行残差相加,从而得到最终的图像特征 h_{new} ,如式(9)所示:

$$h_{new} = h + \lambda_2 h_3 \quad (9)$$

其中, $h_{new} \in \mathbb{R}^{C \times H \times W}$, λ_2 表示网络自动学习并优化的权重系数。

2.5 损失函数

FFGAN 模型的损失函数由生成器损失函数以及鉴别器损失函数两部分组成。其中生成器损失函数包括对抗性损失、DAMSM 损失以及感知损失三部分。生成器损失函数用式(10)表示:

$$L_G = L_{adv}^G + \lambda_{DA} L_{DAMSM} + \lambda_{PL} L_{PL} \quad (10)$$

其中, L_{adv}^G 表示对抗性损失, L_{DAMSM} 表示深度注意力多模态相似损失, L_{PL} 表示感知损失, λ_{DA} , λ_{PL} 分别表示 DAMSM 损失和感知损失所对应的权重。 L_{adv}^G 和 L_{DAMSM} 可以用式(11)、式(12)表示:

$$L_{adv}^G = -\mathbb{E}_{x \sim P_G} [D(\hat{x}, s)] \quad (11)$$

$$L_{DAMSM} = -\sum_{i=1}^M [\log P(w_i | \hat{x}_i) + \log P(\hat{x}_i | w_i) + \log P(s_i | \hat{x}_i) + \log P(\hat{x}_i | s_i)] \quad (12)$$

其中, M 表示句子的数量, $\hat{x}_i$ 表示所生成的图像特征, w_i 表示单词特征, s_i 表示句子特征。

在风格迁移^[13]、图像修复^[14]、超分辨率重建^[15]等任务中,为了减小生成图像和原始图像在像素空间上的差异,有些方法会选择引入感知损失来引导模型的训练。由于 Resnet 网络模型由大量的残差块构成,其能够保留更多的特征信息,并且不会随着网路层数的增加而丢失,故采用预训练好的 Resnet50^[16] 网络提取目标图像和生成图像的深度特征,再计算这两个深度特征的欧式距离,并将其作为感知损失引导网络的训练。感知损失 L_{PL} 如式(13)所示:

$$L_{PL} = \mathbb{E}_{B \sim P(B), \hat{B} \sim P(\hat{B})} \|E_R(\hat{x}) - E_R(x)\|_2 \quad (13)$$

其中, E_R 表示 Resnet50 预训练模型, $\hat{x}$ 表示生成图像特征, x 表示真实图像特征。

鉴别器:采用 DF-GAN^[8] 中的鉴别器损失函数。由式(14)所示:

$$\begin{aligned} L_D = & \mathbb{E}_{x \sim P_{data}} [\max(0, 1 - D(x, s))] \\ & + \frac{1}{2} \mathbb{E}_{x \sim P_G} [\max(0, 1 + D(\hat{x}, s))] \\ & + \frac{1}{2} \mathbb{E}_{x \sim P_{data}} [\max(0, 1 + D(x, \hat{s}))] \\ & + \lambda_{MA} \mathbb{E}_{x \sim P_{data}} [(\|\nabla_x D(x, s)\|_2 + \|\nabla_s D(x, s)\|_2)^p] \end{aligned} \quad (14)$$

其中, x 为目标图像特征, s 表示与目标图像特征 x 对应的文本描述。 $\hat{x}$ 为生成图像特征, $\hat{s}$ 表示不匹配的文本描述。 λ_{MA} 和 p 为 MA-GP 损失的超参数。

3 实验结果与分析

3.1 实验环境和数据集

在本文中,采用SSAGAN模型作为基础模型。训练过程中使用Adam优化器,其中 $\beta_1=0, \beta_2=0.9$ 。生成器和鉴别器的学习率分别设置为0.000 1和0.000 4, batch_size设置为32,超参数 $\lambda_{PL}=0.000\ 05$ 。

本文使用了CUB^[17]以及COCO^[18]数据集来验证所提出模型FFGAN的优越性。CUB数据集包括了11 788张鸟类的图片,其中训练集包含了8 855张图片,测试集含有2 933张图片,一共有200个鸟类类别,并且每张图片都有10条不同的文本描述与之对应。COCO数据集是一个超大规模的数据集,它包含了82 783张图片作为训练集,40 470张图片作为测试集,其中每一张图片对应5条不同的文本描述。CUB数据集和COCO数据的介绍如表1所示。

表1 数据集参数

数据集	训练集数量	测试集数量
CUB	8 855	2 933
COCO	82 783	40 470

3.2 评价指标

本文采用初始分数(Inception Score, IS)^[19]、弗雷歇距离(Fréchet Inception Distance, FID)^[20]作为评价指标。IS是用来评价生成图像的清晰度以及生成图像的多样性的,IS值越高说明生成图像更加清晰并且多样性更好;FID用来衡量生成图像和目标图像之间的距离,即生成图像与目标图像之间的相似程度,FID分数越小,说明生成图像和目标图像越接近,逼真度更高。

3.3 实验结果

3.3.1 定量分析

为了验证本文提出的FFGAN模型的性能,先从定量角度对模型进行了分析。本文在CUB数据集以及COCO数据集上分别随机生成30 000图片。

计算模型的IS分数以及FID分数,并和AttnGAN^[6]、DMGAN^[7]、DAE-GAN^[21]、DF-GAN^[8]和SSAGAN^[9]进行了定量对比实验,其结果表如表2所示。

从表2第二列可知,在CUB数据集上,相较于传统的三阶段生成网络DMGAN^[7],IS分数从4.75提升到了5.22,提升了9.9%。其次,对于单阶段生成网络DF-GAN^[8]和SSAGAN^[9],其IS分数也分别从5.10和5.17提升到了5.22,分别提升了2.3%,1%。结果表

表2 不同模型在CUB和COCO数据集上的性能比较

模型	CUB		COCO
	IS	FID	FID
AttnGAN ^[6]	4.36±0.03	23.98	35.49
DMGAN ^[7]	4.75±0.07	16.09	32.64
DAE-GAN ^[21]	4.42±0.04	15.19	28.12
DF-GAN ^[8]	5.10	14.81	21.42
SSAGAN ^[9]	5.17±0.08	15.61	19.37
FFGAN	5.22±0.08	13.91	16.97

明,本文提出的FFGAN模型相比于现有模型,能够更好地提高生成图像的清晰度。

表2第三列表明了不同模型在CUB数据集上的FID分数。对于DMGAN^[7],FID分数从16.09下降到了13.91。并且相较于DF-GAN^[8]和SSAGAN^[9],FFGAN的FID也有了明显的下降,说明所提出的FFGAN模型生成的图像更接近于真实图像。

由于已有研究表明^[22],IS无法很好地评估COCO数据集上的图像质量,所以本文选用FID分数对FFGAN模型的性能进行评估,其结果如表2第四列所示。从结果中可以看出,FFGAN的FID分数为16.97,相较于DFGAN^[8]以及SSAGAN^[9],分别下降了4.45以及2.4。其也表明了FFGAN模型生成的图像和真实图像的逼真度高。

从上述对比实验可知,本文所提出的FFGAN模型比其他模型生成的图片的清晰度更高,逼真度更好。

3.3.2 定性分析

为了能够直观地展示FFGAN模型的优越性,本文将FFGAN生成的图片和DMGAN^[7]、SSAGAN^[9]生成的图片进行了对比分析,其对比图如图4和图5所示。

图4表明了不同模型在CUB数据集上进行的图片结果展示。其中第一行显示了生成图像所对应的输入文本,第二行到第四行分别展示了DMGAN^[7]、SSAGAN^[9]以及FFGAN所对应文本生成的图片。从图中可以看出,虽然DMGAN^[7]能够生成一张较为完整的鸟类图片,但是图片较为模糊并且一些细节信息也不够清晰,比如鸟类的脚等;SSAGN^[9]在DMGAN^[7]的基础上进一步提高了图片的纹理信息,但是其细节提升能力还不够;本文的FFGAN相较于DMGAN^[7]、SSAGAN^[9]不仅提升了图片的清晰度,而且还使生成的图片具有更加丰富的细节。

3.3.3 消融实验

为了验证本文所提出的基于特征融合的文本到图像生成方法的有效性,本文在CUB数据集上进行了



图4 CUB数据集上的生成图片比较



图5 COCO数据集上的生成图片比较

消融实验,结果如表3所示。首先本文在基线模型SSAGAN^[9]中添加感知损失模块,结果表明IS分数从5.07增加到了5.13, FID分数从15.61降低到了14.37,说明加入感知损失模块之后,能够有效提升模型生成图片的清晰度以及逼真度;然后在此基础上引入文本-图像跨模态融合模块, FID分数从14.37降低到了13.54,说明文本-图像跨模态融合模块能够有效捕捉文本信息,从而使生成的图像特征更加符合真实图像的特征分布。

表3 在CUB数据集上的消融实验

方法	IS	FID
基线模型 ^[9]	5.07±0.04	15.61
基线模型+感知损失	5.13±0.06	14.37
FFGAN	5.10±0.04	13.54

在SSAGAN^[9]模型中,文本编码器和图像编码器也参与了网络的训练。在本文中也基于此方法对模型进行训练。结果如表4所示。从表4中可以得知:

较于基线模型SSAGAN^[9], IS分数从5.17增加到了5.22, FID分数从16.58下降到了13.91,体现出本文所提模型FFGAN的优越性。

表4 基于文本编码器、图像编码器训练的消融实验

方法	IS	FID
基线模型 ^[9]	5.17±0.08	16.58
FFGAN	5.22±0.08	13.91

4 结论

本文提出了一种基于特征融合的文本来到图像生成方法。本文所提出的文本-图像跨模态融合模块能够充分获得深层次的文本信息,从而使得生成的图片具有更加完善的细节信息,同时引入感知损失进一步细化生成的图片,使得生成图像特征和目标图像特征更加接近,从而生成清晰度更好,细节特征更加丰富以及逼真度更高的图片。虽然大量的实验对比结果表明本文所提方法的有效性,但是本文所设计的文本-图像跨模态融合模块还需要更加完善,在接下来的工

作中,会进一步对文本-图像融合模块进行改进,从而使生成图片能够更好地表征文本描述。

参考文献

- [1] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[J]. Communications of the ACM, 2020, 63(11): 139-144.
- [2] PATHAK D, KRÄHENBÜHL P, DONAHUE J, et al. Context encoders: Feature learning by inpainting[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 2536-2544.
- [3] ANDREINI P, BONECHI S, BIANCHINI M, et al. Image generation by GAN and style transfer for agar plate image segmentation [J]. Computer Methods and Programs in Biomedicine, 2020, 184: 105268.
- [4] REED S, AKATA Z, YAN X C, et al. Generative adversarial text to image synthesis[C]//Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48. New York: ACM, 2016: 1060-1069.
- [5] ZHANG H, XU T, LI H S, et al. StackGAN: text to photo-realistic image synthesis with stacked generative adversarial networks [C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2017: 5908-5916.
- [6] XU T, ZHANG P C, HUANG Q Y, et al. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 1316-1324.
- [7] CHEN Z L, WANG C, WU H M, et al. DMGAN: Discriminative metric-based generative adversarial networks[J]. Knowledge-Based Systems, 2020, 192: 105370.
- [8] TAO M, TANG H, WU F, et al. DF-GAN: A simple and effective baseline for text-to-image synthesis[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 16494-16504.
- [9] LIAO W T, HU K, YANG M Y, et al. Text to image generation with semantic-spatial aware GAN[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2022: 18166-18175.
- [10] SAHARIA C, CHAN W, SAXENA S, et al. Photorealistic text-to-image diffusion models with deep language understanding[J]. Advances in Neural Information Processing Systems, 2022, 35: 36479-36494.
- [11] RAMESH A, DHARIWAL P, NICHOL A, et al. Hierarchical text-conditional image generation with CLIP latents[EB/OL]. [2023-8-15]. <https://3dvar.com/Ramesh2022Hierarchical.pdf>.
- [12] 赖丽娜, 米瑜, 周龙龙, 等. 生成对抗网络与文本图像生成方法综述[J/OL]. 计算机工程与应用, 2023: 1-23. (2023-03-15) [2023-09-23]. <https://kns.cnki.net/kcms/detail/11.2127.TP.20230314.1549.022.html>.
- [13] WANG L, WANG L, CHEN S. ESA-CycleGAN: Edge feature and self-attention based cycle-consistent generative adversarial network for style transfer[J]. IET Image Processing, 2022, 16(1): 176-190.
- [14] LUO X, HE X, QING L, et al. EyesGAN: Synthesize human face from human eyes[J]. Neurocomputing, 2020, 404: 213-226.
- [15] SONG J, YI H W, XU W Q, et al. Dual perceptual loss for single image super-resolution using ESRGAN[EB/OL]. [2023-8-15]. <https://arxiv.org/pdf/2201.06383.pdf>.
- [16] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 770-778.
- [17] WAH C, BRANSON S, WELINDER P, et al. The Caltech-UCSD Birds-200-2011 Dataset[R]. Pasadena, USA: California Institute of Technology, Computation & Neural Systems Technical Report, 2011: CNSTR-2010-001.
- [18] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: Common objects in context[C]//Computer Vision — ECCV 2014. Cham: Springer International Publishing, 2014: 740-755.
- [19] SALIMANS T, GOODFELLOW I, ZAREMBA W, et al. Improved techniques for training GANs[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems. New York: ACM, 2016: 2234-2242.
- [20] HEUSEL M, RAMSAUER H, UNTERTHINER T, et al. GANs trained by a two time-scale update rule converge to a local Nash equilibrium[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 6629-6640.
- [21] RUAN S L, ZHANG Y, ZHANG K, et al. DAE-GAN: Dynamic aspect-aware GAN for text-to-image synthesis[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2022: 13940-13949.
- [22] ZHANG Z X, SCHOMAKER L. DTGAN: dual attention generative adversarial networks for text-to-image generation[C]//2021 International Joint Conference on Neural Networks (IJCNN). Piscataway: IEEE, 2021: 1-8.

作者简介



周 杭 (1999—), 男, 重庆人, 硕士研究生。研究方向: 图像生成。

何小海 (1964—), 男, 四川绵阳人, 教授, 现为四川大学博士生导师。研究方向: 图像处理, 模式识别, 图像通信。

卿 斌 波 (1982—), 男, 四川简阳人, 教授, 现为四川大学博士生导师。研究方向: 图像处理, 模式识别, 视频通信。