计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210600085

基于 Transformer 交叉注意力的文本生成图像技术

谈馨悦　何小海　王正勇　罗晓东　卿粼波
四川大学电子信息学院　成都 610065
(2019222055241@stu.scu.edu.cn)

摘　要　近年来,以生成对抗网络为基础的从文本生成图像方法的研究取得了一定的进展。文本生成图像技术的关键在于构建文本信息和视觉信息间的桥梁,促进网络模型生成与对应文本描述一致的逼真图像。目前,主流的方法是通过预训练文本编码器来完成对输入文本描述的编码,但这些方法在文本编码器中未考虑与对应图像的语义对齐问题,独立对输入文本进行编码,忽略了语言空间与图像空间之间的语义鸿沟问题。为解决这一问题,文中设计了一种基于交叉注意力编码器的对抗生成网络(CAE-GAN),该网络通过交叉注意力编码器,将文本信息与视觉信息进行翻译和对齐,以捕捉文本与图像信息之间的跨模态映射关系,从而提升生成图像的逼真度和与输入文本描述的匹配度。实验结果表明,在 CUB 和 coco 数据集上,与当前主流的方法 DM-GAN 模型相比,CAE-GAN 模型的 IS(Inception Score)分数分别提升了 2.53%和 1.54%,FID (Fréchet Inception Distance)分数分别降低了 15.10%和 5.54%,由此可知,CAE-GAN 模型生成图像的细节更加完整、质量更高。

关键词:文本描述生成图像;生成对抗网络;交叉注意力编码;图像生成;计算机视觉
中图法分类号　TP183

Text-to-Image Generation Technology Based on Transformer Cross Attention

TAN Xin-yue, HE Xiao-hai, WANG Zheng-yong, LUO Xiao-dong and QING Lin-bo
College of Electronics and Information Engineering, Sichuan University, Chengdu 610065, China

Abstract　In recent years, the research on the methods of text to image based on generative adversarial network (GAN) continues to grow in popularity and have made some progress. The key of text-to-image generation technology is to build a bridge between the text information and the visual information, and promote the model to generate realistic images consistent with the corresponding text description. At present, the mainstream method is to complete the encoding of the descriptions of the input text by pre-training the text encoder, but these methods do not consider the semantic alignment with the corresponding image in the text encoder, and adopt the independent encoding of the input text, ignoring the semantic gap between the language space and the image space. To address the problem, in this paper, a generative adversarial network based on the cross-attention encoder (CAE-GAN) is proposed. The network uses a cross-attention encoder to translate and align text information with visual information, and captures the cross-modal mapping relationship between text and image information, so as to improve the fidelity of the generated images and the matching degree with input text description. The experimental results show that, compared with the DM-GAN model, the inception score (IS) of CAE-GAN model increases by 2.53% and 1.54% on CUB and coco datasets, respectively. The fréchet inception distance score decreases by 15.10% and 5.54%, respectively, indicating that the details and the quality of the images generated by the CAE-GAN model are more perfect.

Keywords　Text-to-Image generation, Generative adversarial networks, Cross-attention encoding, Image generation, Computer vision

---

到稿日期:2021-06-08　返修日期:2021-10-20
基金项目:国家自然科学基金(61871278, U1836118);成都市重大科技应用示范项目(2019-YF09-00120-SN);四川省科技计划项目(2018HH0143)
This work was supported by the National Natural Science Foundation of China(61871278, U1836118), Chengdu Major Technology Application Demonstration Project(2019-YF09-00120-SN) and Sichuan Science and Technology Program(2018HH0143).
通信作者:何小海(hxh@scu.edu.cn)

1 引言

近年来，基于文本和图像跨模态融合的研究吸引了计算机视觉和自然语言处理领域学者的广泛关注，尤其是在图像描述生成[1-2]、视觉问答[3]、视觉推理[4]以及从文本生成图像[5]等方面，例如，Xu 等引入了一个基于注意力的模型，通过自动学习图像内容描述[6]，实现了基于卷积神经网络视觉特征的跨模态检索；Wei 等提出了一种深度语义匹配的方法[7]用于解决一个或多个标签标注的样本图片跨模态检索问题；Bi 等[8]提出了一种以谓语动词为结构中心的句法分析方法，用来解决图像描述生成中图像内容分析与生成文本的过程缺乏结构信息的问题；Chen 等[9]提出了一个关于目标像素点非对称的区域作为目标图像块的非对称非局部变分模型，进一步提升图像的复原效果；Xu 等[10]提出了一种基于跨模态生成对抗网络的甲状腺超声图像与文本报告互检的方法，将图像文本跨模态研究应用在医疗领域，同时这些研究在军事、农业、安防、教育等领域也都具有广泛的应用价值。本文聚焦于文本图像跨模态融合领域中热度较高的子任务，即从文本描述生成图像。

目前，对从文本生成图像的方法的研究取得了很大的进步和成果，经典模型包括 GAN-INT-CLS 模型[11]、StackGAN 模型[12]和 AttnGAN 模型[13]等，在这些经典模型的基础上，研究者们针对生成图像结构局部扭曲和纹理不清晰的问题，提出了 MLGAN 模型[14]、MPRGAN 模型[15]和结合互信息最大化的文本生成图像模型[16]等。Huang 等[17]还针对生成图像局部对象重叠和缺失的问题，提出了一种基于场景描述的文本生成图像网络模型，然而现有的方法仍然存在一些可待优化的地方。由于文本语言和视觉图像是两种不同模态的信息，如何在两种模态信息之间建立起正确的沟通桥梁是文本生成图像任务的关键，生成图像的质量也取决于该沟通机制和相应算法的优劣。当前主流的方法是采用多级生成对抗网络，该方法生成图像的质量高度依赖于初始级生成图像的质量，而初始级生成图像的质量又与文本描述的语义息息相关。目前的网络模型在文本编码阶段未考虑文本信息和图像信息的对应相关性，初始级生成图像与文本语义的匹配度不高，同时细节表现也较差，导致后续优化、精炼图像细节的难度较大而影响到生成的图像质量。针对上述问题，本文提出了一种基于交叉注意力编码的从文本生成图像的方法，其特点是，在编码阶段将文本表达和视觉表达进行对应关系的构建，从而提升生成图像与文本描述的匹配度，并提高生成图像的质量和多样性。

本文的主要贡献如下：

(1)在编码阶段，引入了一种基于 Transformer 的交叉注意力机制，将文本表示与图像表示进行对齐和翻译，挖掘文本与图像之间跨模态的映射关系，并针对不同的任务进行预训练，将视觉信息和语言信息有效联系，从而改善生成图片的质量。

(2)设计了一种基于 Transformer 交叉注意力编码的生成对抗网络(Generative Adversarial Network Based on the Cross-Attention Encoder，CAE-GAN)，该网络能够有效地捕捉图像与文本之间的内在联系，在文本描述生成图像的任务上，定量和定性的评价结果都取得了较大提升。

在 CUB[18]和 coco[19]这两个数据集上对本文设计的网络模型 CAE-GAN 进行训练和验证，实验结果表明，CAE-GAN 模型的性能得到了显著提升，评价指标 IS 和 FID 上也得到了显著提升。

2 相关工作

2.1 从文本生成图像

文本生成图像属于计算机视觉和自然语言处理交叉领域，对其进行研究具有较大的挑战性。文本生成图像，即根据文本描述生成符合文本内容和涵盖丰富细节的高质量逼真图像。由于同一段描述性文本所生成的图像可能不完全一致，因此此任务生成的图像也具备多样性。

随着深度学习尤其是生成对抗网络(Generative Adversarial Networks，GANs)[20]的发展，其在解决跨模态任务上展示出了可观的效果。近年来，生成对抗网络用于图像和视频生成的应用率显著提升，常被用于文本生成图像任务，且展现出了广阔的应用前景，为人工智能的发展作出了进一步的贡献。

生成对抗网络(GANs)[20]是由蒙特利尔大学的 Goodfellow 等于 2014 年提出的一个深度学习新框架。生成对抗网络模型由两个重要组件构成：生成器 G 和判别器 D，两者是独立的网络，互相竞争和对抗。在训练过程中，生成器的目标是尽可能地生成逼真图像用于“欺骗”鉴别器，直到鉴别器无法辨认出生成的图像是假图像时达到一种拟合状态，即生成器和判别器的训练实际是一个动态的“博弈”过程。由于生成对抗网络具有一定的随机性，因此，为了使生成过程受到一定条件约束，条件生成对抗网络(Condition Generative Adversarial Networks，cGANs)[21]被提出。其原理是在 GAN 的基础上添加了附加条件作为输入，从而使生成对抗网络能够获取外部信息，生成过程受到附加条件的约束。

大多数研究文本生成图像方法的原理是将文本描述作为生成器的附加条件信息，采用单阶段或多阶段的方式生成逼真图像。前者使用单个生成器和判别器生成图像，后者使用多级生成器和判别器分阶段地逐级生成有更高分辨率且细节更丰富的图像。Reed 等于 2016 年提出了 GAN-INT-CLS 模型[11]，在 Oxford-102[22]和 CUB 数据集上，该模型使用单级生成对抗网络生成了分辨率为 64 * 64 的图像。文献[23]提出的 GAWWN 模型在 GAN-INT-CLS 模型的基础上增加了边界限定和关键点，由此生成了分辨率为 128 * 128 的图像。Zhang 等提出的 StackGAN 模型[12]使用两级生成对抗网络分步生成了分辨率为 256 * 256 的图像。此后，为了进一步提

升图像精度和改善图像细节，Zhang 等提出了 StackGAN++模型[24]。该模型采用三级生成对抗网络逐级生成分辨率为 64 * 64，128 * 128，256 * 256 的图像，实现了图像从低分辨率到高分辨率的分步生成，减少了信息残缺的问题，从而改善了图像细节，提升了图像质量。在后续研究中，模型 AttnGAN[13]添加了注意力机制，将注意力精确到了单词级别，将提取的文本句子信息和单词信息作为局部约束输入到生成器中，从而优化了图像的细节。同时，还新增了 DAMSM[13](Deep Attentional Multimodal Similarity Model)模块，通过改进训练过程中计算损失函数的方式，使生成图像的质量得到了进一步的提升。为了解决生成图像质量过度依赖于初始生成图像的问题，Zhu 等提出了一个动态记忆生成对抗网络(Dynamic Memory Generative Adversarial Networks，DM-GAN)[25]，该模型提出了一种动态存储模块，通过动态更新各级生成图像的特征，在一定程度上降低了最终生成图像质量对初始生成图像质量的依赖度，且丰富了图像细节。但是，目前的方法在对文本描述进行编码时，并未考虑与对应图像的映射关系，导致初始阶段生成图像与文本语义的匹配度仍然较低，且图像质量也受到了影响。为此本文在训练文本编码器时引入交叉注意力编码的模式，让文本编码器学习语言空间与图像空间的映射关系，从而更准确地对文本描述进行编码，促进生成网络生成更逼真的图像。

2.2 交叉注意力

为提升生成图片的质量，文献[26]将自注意力机制引入生成对抗网络，在一定程度上提升了生成图像的清晰度。Ju 等[27]也在 AttnGAN 的基础上，将自注意力机制应用到初始图像生成阶段，并将文本特征细化成句子特征和单词特征，从而提升了模型的稳定性。然而，文本生成图像是视觉语言推理跨模态研究的子问题，对于视觉和语言推理问题而言，理解视觉概念和语言语义是十分重要的，尤其是学习这两者之间的联系和对齐。最近，基于 Transformer[28]来学习跨模态的编码表示的研究取得了重大进展，这种跨模态的交叉注意力机制能有效提升视觉和语言相关任务的模型性能。Li 等提出了 Unicoder-VL 模型[29]，即以预训练的方式学习视觉和语言联合表示的通用编码器，该模型在图像文本检索和视觉常识推理任务中都取得了显著的成果。Wang 等于 2019 年提出了 CAMP(Cross-modal Adaptive Message Passing for Text-image Retrieval)模型[30]，该模型可以自适应地控制跨模态消息传递的信息流，在文本图像跨模态检索任务中得到了明显的改善。一些广泛用于建模视觉和语言任务的模型框架，如 VisualBERT[31]，ViLBERT[32]和 LXMERT[33]等，均旨在捕捉语言和视觉的内在联系，通过有效地搭建视觉信息和语言信息的沟通桥梁，来解决常见的视觉语言交叉子任务。就文本合成图像任务而言，文本描述和原图像分别代表了语言语义信息和图像视觉信息，分别提取文本描述和原图像的特征，若在编码阶段将他们通过类似的交叉注意力机制联系起来，捕捉文本描述信息与图像信息之间的映射关系，则可能会优化生成图像的细节，提高图像质量。受此启发，本文提出了基于交叉注意力编码的生成对抗网络(CAE-GAN)。

3 基于 Transformer 交叉注意力的生成对抗网络

如图 1 所示，本文设计的基于 Transformer 交叉注意力的生成对抗网络(CAE-GAN)的整体架构由 3 部分构成，即预训练编码器、动态存储模块和三级对抗生成网络。其中，预训练编码器是本文引入的交叉注意力机制中用于训练文本的编码器，其原理是将文本描述输入到交叉注意力编码器(cross attention encoder)中，该编码器会输出一个与图像特征进行对齐和翻译后的交叉注意力特征向量 $\boldsymbol{f}_c$，以及一个单词特征矩阵 $\boldsymbol{W}$。交叉注意力特征向量 $\boldsymbol{f}_c$ 与随机采样的高斯噪声相结合，作为经典三级生成对抗网络的输入，由三级生成器逐级生成与文本描述相符合的高质量图像。单词特征矩阵 $\boldsymbol{W}$ 输入到动态存储模块[25]与初级图像特征进行融合，得到融合后的新图像特征。上述过程如下：

$$\boldsymbol{f}_c,\boldsymbol{W}=C_E(\boldsymbol{s},\boldsymbol{F}_R) \tag{1}$$

$$\boldsymbol{F}_0=G_0(\boldsymbol{f}_c+\boldsymbol{z}) \tag{2}$$

$$\boldsymbol{F}_1=G_1(DM(\boldsymbol{F}_0,\boldsymbol{W})) \tag{3}$$

$$\boldsymbol{F}_2=G_2(DM(\boldsymbol{F}_1,\boldsymbol{W})) \tag{4}$$

其中，C_E 表示交叉注意力编码器，DM 表示动态存储模块，G_0，G_1，G_2 分别表示三级生成器，$\boldsymbol{F}_R$ 表示原图像特征，$\boldsymbol{s}$ 是从文本描述中提取出的全局句子向量，$\boldsymbol{z}$ 表示随机噪声，$\boldsymbol{F}_0$，$\boldsymbol{F}_1$，$\boldsymbol{F}_2$ 分别表示 G_0，G_1，G_2 生成图像的特征。

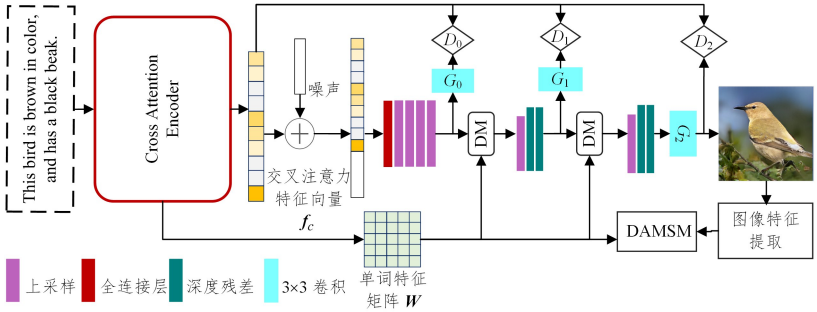

图 1 CAE-GAN 模型框架图

Fig. 1 CAE-GAN model frame diagram

3.1 交叉注意力编码

根据文本描述生成图像需要理解描述语句和图像之间的对应关系，其关键在于解决图像空间和文本空间的语义鸿沟。本文设计了一种交叉注意力编码器，对语言信息和视觉信息进行联合交叉编码和对齐。

如图 2 所示，交叉注意力编码器由 4 部分组成，包括文本特征提取、图像特征提取、交叉注意力编码和自注意力编码。

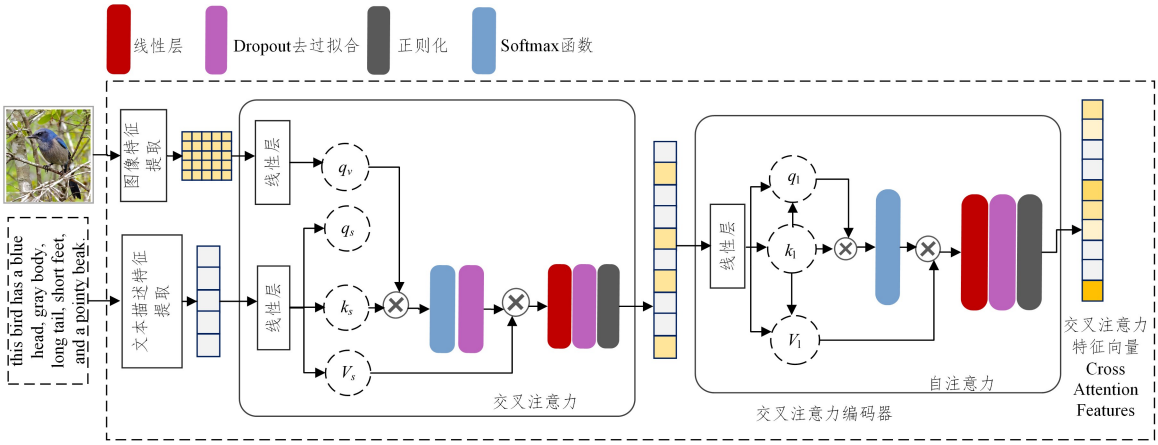

图 2 交叉注意力编码器框架图

Fig. 2 Cross attention encoder diagram

(1)文本特征提取。文本特征提取部分以双向长短时记忆(LSTM)[34]网络为基础，其目的是将原始的文本描述编码，输出单词特征矩阵 $\boldsymbol{W}^{d\times t}$ 和一个全局句子特征向量 $\boldsymbol{s}$，$\boldsymbol{s}$ 包含了文本描述的语义信息。在双向 LSTM 中，每一个单词特征对应了每一个隐藏层，包括了两个方向；而全局的句子特征是由最后一个隐藏层的两个方向连接得出，包含了整个句子描述的语言语义信息。

$$\boldsymbol{W}^{d\times t},\boldsymbol{s}=T_E(T_{\text{text}}) \tag{5}$$

其中，d 表示词向量的维度；t 表示单词的个数；T_E 表示一个双向 LSTM 网络，用于提取文本特征；T_{text} 表示原文本描述。

(2)图像特征提取。图像特征提取部分使用 InceptionV3 模型[35]来提取原图像的特征。InceptionV3 网络的本质为卷积神经网络(CNN)，卷积神经网络的中间层学习图像不同子区域的局部特征，深层学习图像的全局特征。

$$\boldsymbol{f}_v=InC(\boldsymbol{I}_{\text{img}}) \tag{6}$$

(3)交叉注意力编码。交叉注意力编码部分主要用于交换信息，并构建语言特征与图像特征的内部联系，实现两者的联合编码。全局句子特征 $\boldsymbol{s}$ 和图像特征 $\boldsymbol{f}_v$ 经交叉注意力编码后，融合成一个新的特征 $\boldsymbol{l}_c$，$\boldsymbol{l}_c$ 是语言信息和视觉信息的联合编码。文本特征 $\boldsymbol{s}$ 和图像特征 $\boldsymbol{f}_v$ 经过线性层映射到两个特征空间，分别是 $\boldsymbol{q}_s,\boldsymbol{k}_s,\boldsymbol{v}_s$ 和 $\boldsymbol{q}_v,\boldsymbol{k}_v,\boldsymbol{v}_v$。

$$\boldsymbol{q}_s,\boldsymbol{k}_s,\boldsymbol{v}_s=Linear(\boldsymbol{s}) \tag{7}$$

$$\boldsymbol{q}_v,\boldsymbol{k}_v,\boldsymbol{v}_v=Linear(\boldsymbol{f}_v) \tag{8}$$

利用 $\boldsymbol{q}_s,\boldsymbol{k}_s,\boldsymbol{v}_s$ 和 $\boldsymbol{q}_v,\boldsymbol{k}_v,\boldsymbol{v}_v$ 计算交叉注意力分数，且利用 Softmax 函数和 Dropout 函数进行归一化和防止过拟合处理，整个计算过程如下：

$$score=\lambda_c\,\boldsymbol{q}_v\,\boldsymbol{k}_s^{\mathrm{T}} \tag{9}$$

$$score'=Softm(score) \tag{10}$$

$$\boldsymbol{s}_c=dropout(score') \tag{11}$$

$$\boldsymbol{l}=\boldsymbol{s}_c\cdot\boldsymbol{v}_s \tag{12}$$

其中，λ_c 为自定义常数，$Softm$ 表示 softmax 函数。

将 $\boldsymbol{l}$ 经过全连接层和规范化处理，得到 $\boldsymbol{l}_c$。

$$\boldsymbol{l}_c=Normalization(A_1\boldsymbol{l}+\boldsymbol{B}_1) \tag{13}$$

其中，A_1，B_1 是待学习的参数。

(4)自注意力编码。为了进一步检索上下文信息，将交叉注意力编码部分的输出 $\boldsymbol{l}_c$ 输入到自注意力编码部分。首先，$\boldsymbol{l}_c$ 通过线性层映射到一个特征空间 $\boldsymbol{q}_l,\boldsymbol{k}_l,\boldsymbol{v}_l$，并通过 $\boldsymbol{q}_l,\boldsymbol{k}_l,\boldsymbol{v}_l$ 来计算自注意力权重；然后，$\boldsymbol{l}_c$ 经自注意力部分处理得到交叉注意力编码特征向量 $\boldsymbol{f}_c$，$\boldsymbol{f}_c$ 会与高斯噪声加性结合后送入生成对抗网络；最后，由三级生成器逐级生成与文本描述对应的逼真图像。上述过程可表示为：

$$\boldsymbol{q}_l,\boldsymbol{k}_l,\boldsymbol{v}_l=Linear(\boldsymbol{l}_c) \tag{14}$$

$$\boldsymbol{s}_s=Dropout(softm(\lambda_s\,\boldsymbol{q}_l\,\boldsymbol{k}_l^{\mathrm{T}})) \tag{15}$$

$$\boldsymbol{l}_{cs}=\boldsymbol{s}_s\cdot\boldsymbol{v}_l \tag{16}$$

$$\boldsymbol{f}_c=Normalization(A_2\,\boldsymbol{l}_{cs}+B_2) \tag{17}$$

其中，A_2，B_2 为需要学习的参数，$\boldsymbol{f}_c$ 为最终输出的交叉注意力特征向量。

3.2 经典三级生成对抗网络

如图 1 所示，CAE-GAN 模型利用了与 StackGAN[12]，AttnGAN[13]，DM-GAN[25]相似的经典三级对抗生成网络结构，由 G_0，G_1 和 G_2 逐级生成分辨率为 64×64，128×128，256×256 的图片。G_0 在第一阶段生成低分辨率的初始图像，包含了 1 个全连接层、3 个上采样层以及 1 个大小为 $3*3$ 的卷积网络层；G_1 和 G_2 在第二阶段进一步细化初始图像，它们具有相同的结构，包括 1 个上采样层、2 个深度残差网络层以及 1 个大小为 $3*3$ 的卷积网络层。

3.3 动态存储模块

生成器 G_0 与生成器 G_1、生成器 G_1 与生成器 G_2 之间均通过动态存储模块[25]连接，该模块在细化图像阶段进一步融合了图像特征和文本特征。如图 3 所示，动态存储模块由内存写入、键寻址、值读取、响应这 4 步来实现。

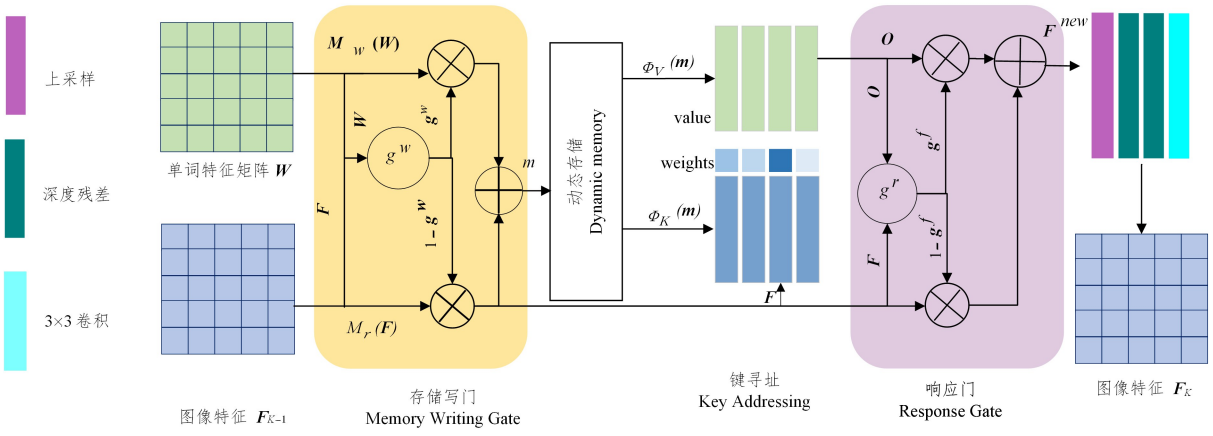

图 3 动态存储模块框图

Fig. 3 Dynamic memory module diagram

该模块的输入为：

$$\boldsymbol{W}=\{\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_T\},\boldsymbol{w}_i\in\mathbb{R}^{N_W} \tag{18}$$

$$\boldsymbol{F}_i=\{\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_N\},\boldsymbol{f}_i\in\mathbb{R}^{N_f},i=0,1 \tag{19}$$

其中，$\boldsymbol{W}$ 为单词特征矩阵；$\boldsymbol{F}_i$ 表示图像特征，如：$\boldsymbol{F}_0$ 为初始生成图像特征，$\boldsymbol{F}_1$ 为第二级生成图像特征等；T 代表单词个数；N_W 为每个单词特征向量的维度；N 表示图像像素数目；N_f 为图像像素特征向量的维度。

内存写入步骤通过存储写门实现，选择相关单词信息优化图像细节：

$$g_i^w(\boldsymbol{F},w_i)=\sigma\left(\boldsymbol{A}*w_i+\boldsymbol{B}*\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{f}_i\right) \tag{20}$$

$$\boldsymbol{m}_i=M_w(\boldsymbol{w}_i)*\boldsymbol{g}_i^w+M_f\left(\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{f}_i\right)*(1-\boldsymbol{g}_i^w) \tag{21}$$

其中，σ 表示 sigmoid 函数，$\boldsymbol{A}$ 和 $\boldsymbol{B}$ 分别表示维度为 $1\times N_w$ 和 $1\times N_f$ 的矩阵，$M_w(\cdot)$和 $M_f(\cdot)$均为 1×1 卷积操作，目的是将单词特征和图像特征映射到 N_m 维度的特征空间。

在键寻址步骤中，该模块计算每个图像特征和每个存储单元特征之间的权重：

$$\alpha_{i,j}=\frac{\exp(\Phi_K(\boldsymbol{m}_i)^{\mathrm{T}}\boldsymbol{f}_j)}{\sum_{i=1}^{T}\exp(\Phi_K(\boldsymbol{m}_i)^{\mathrm{T}}\boldsymbol{f}_j)} \tag{22}$$

其中，$\alpha_{i,j}$表示第 i 个存储单元与第 j 个图像特征之间的相似概率，$\Phi_K(\cdot)$表示 1×1 卷积网络，其目的是映射特征到 N_f 维度。

在值读取步骤中，该模块输出相应的存储表示向量如下。

$$\boldsymbol{o}_j=\sum_{i=1}^{T}\alpha_{i,j}\Phi_V(\boldsymbol{m}_i) \tag{23}$$

响应步骤通过响应门实现，响应门可动态控制信息流动和及时更新图像特征。

$$g_i^f=\sigma(\boldsymbol{W}[\boldsymbol{o}_i,\boldsymbol{f}_i]+b) \tag{24}$$

$$\boldsymbol{f}_i^{\mathrm{new}}=\boldsymbol{o}_i*\boldsymbol{g}_i^f+\boldsymbol{f}_i*(1-\boldsymbol{g}_i^f) \tag{25}$$

其中，$\boldsymbol{g}_i^f$ 表示融合信息的响应门，σ 表示 sigmoid 函数，$\boldsymbol{W}$ 和 b 是参数矩阵和偏置项。

3.4 损失函数

CAE-GAN 模型的损失函数由生成器损失函数和判别器损失函数两部分构成。

生成器损失函数 L 分为以下 3 部分：即条件损失函数 L_{CA}、生成损失函数 L_{G_i} 以及深度注意力多模态相似模型(DAMSM)损失函数 L_{DAMSM}[13]。

$$L=\sum_i L_{G_i}+\tau_1 L_{CA}+\tau_2 L_{DAMSM} \tag{26}$$

其中，τ_1 和 τ_2 分别是 L_{CA} 和 L_{DAMSM} 的权重，L_{G_i} 和 L_{CA} 分别为：

$$L_{G_i}=-\frac{1}{2}[E_{x\sim pG_i}\log D_i(x)+E_{x\sim pG_i}\log D_i(x,s)] \tag{27}$$

$$L_{CA}=D_{KL}(N(\mu(s))\parallel N(0,I)) \tag{28}$$

式(2)中，前一项为非条件损失，用于使生成图像更加逼真；后一项为条件损失，用于使生成图像和文本描述相符合。L_{CA} 则通过源于独立高斯分布的输入句子向量来进行重采样，以避免过拟合，同时也具备了增强训练数据的作用。

DAMSM[13] 损失函数用来测量文本描述与生成图像之间的匹配程度，该损失函数计算了生成图像和整个句子描述匹配的后验概率。通过优化该损失函数来提升生成图像与文本描述的符合程度，并提高图像质量。

判别器损失函数分为条件损失 L_{CD} 和非条件损失 L_D 两部分：

$$L_{D_i}=-\frac{1}{2}[L_D+L_{CD}] \tag{29}$$

$$L_D=E_{x\sim P_{\mathrm{data}}}\log D_i(x)+E_{x\sim pG_i}\log(1-D_i(x)) \tag{30}$$

$$L_{CD}=E_{x\sim P_{\mathrm{data}}}\log D_i(x,s)+E_{x\sim pG_i}\log(1-D_i(x,s)) \tag{31}$$

其中，非条件损失 L_D 用于区别生成图像和真实图像，条件损失 L_{CD} 用于判别生成图像和文本是否相符。

4 实验

4.1 实验数据集

本文从定量和定性的两个方面来评估 CAE-GAN 模型性能。在 CUB 和 coco 两个数据集上进行了训练和测试，数据集的具体情况如表 1 所列。

表 1 数据集

Table 1 Datasets

Datasets	Number of training set images	Number of test set images
CUB	8855	2933
coco	82783	40470

4.2 实验过程

本文实验选取了两个子任务，分别使用 CUB 和 coco 数据集对其进行训练和测试。

实验步骤分为 3 步，即预训练编码器、训练整个模型以及测试整个模型性能。

预训练编码器是本文实验的重心和关键部分，通过不同的子任务来预训练本文设计的交叉注意力编码器，以捕捉不同子任务中文本信息与图像信息的映射关系，从而得到针对该子任务的文本描述并与对应图像信息的联合编码相结合。将预训练编码器这一步骤独立出来，能够使交叉注意力编码器的应用更加灵活，其结果是得到并保存已训练好的编码器模型。

在训练整个模型时，首先需要加载已保存的编码器模型，然后独立训练 CAE-GAN 模型的其他部分。

在测试整个模型阶段，对于 CUB 数据集和 coco 数据集的测试集，本文的 CAE-GAN 模型均生成了 30000 张逼真图片，分别计算其 IS 分数和 FID 分数，并通过这两个指标来定量评价本文提出的 CAE-GAN 模型的性能。

4.3 评价指标

本文用 IS[36] 和 FID 分数[37] 来衡量 CAE-GAN 的性能。AttnGAN 模型[13] 和 DM-GAN 模型[25] 也利用 R-precision 指标来衡量生成图像和文本的匹配程度，其原理是计算文本描述句子向量和全局图像向量的距离。但由于交叉注意力编码器在预训练编码阶段已将全局句子向量和全局图像特征进行了对齐，因此在本实验中此指标不具备衡量的价值。

IS 是评价 GAN 网络性能的一个重要指标，IS 的计算式为：

$$IS=e^{E_{x\in p}D_{KL}(p(y|x)\parallel p(y))} \tag{32}$$

其中，$p(y|x)$ 和 $p(y)$ 分别表示由预训练图像编码器模型预测的标签 y 的条件概率和边缘概率，KL 散度与这两者之间的关系为：若条件概率越低，边缘概率越高，则 KL 散度越大。而 KL 散度值越高，代表图像质量越高，更具多样性。因此，IS 指标越大，说明生成图像质量越高且具有更加丰富的多样性。

FID 分数是用于计算生成图像特征和真实图像特征之间距离的一种度量。FID 值越低意味着两者的特征更加接近，说明生成图像更加接近真实图像，即生成图像更加生动形象。通过图像编码器提取生成图像和真实图像的特征，并根据它们的特征均值和特征协方差得出 FID 分数，其计算式如下：

$$FID=\parallel \mu_\gamma-\mu_g \parallel +T_\gamma\left(\sum_\gamma+\sum_g-2\left(\sum_\gamma\sum_g\right)^{\frac{1}{2}}\right) \tag{33}$$

其中，μ_γ，μ_g，$\sum_\gamma$，$\sum_g$ 分别代表了真实图像均值、生成图像均值、真实图像特征协方差以及生成图像特征协方差。

4.4 实验结果

4.4.1 定量评价

本文从定量和定性两个方面来评估 CAE-GAN 模型的性能。本文在 CUB 数据集和 coco 数据集的测试集中随机生成了 30000 张图片来计算 IS 分数和 FID 分数，并与当前主流的 StackGAN[12]，AttnGAN[13]，DM-GAN[25] 进行比较，定量对比实验的结果如表 2、表 3 所列。

表 2 不同模型在 CUB 数据集上的 IS 和 FID 分数

Table 2 IS and FID scores of different models on CUB dataset

Models	IS	FID
StackGAN[12]	3.70±0.04	35.11
AttnGAN[13]	4.36±0.03	23.98
DM-GAN[25]	4.75±0.07	16.09
SegAttnGAN[38]	4.82±0.05	—
Our CAE-GAN	**4.87±0.06**	**13.66**

表 3 不同模型在 coco 数据集上的 IS 和 FID 分数

Table 3 IS and FID scores of different models on coco dataset

Models	IS	FID
StackGAN[12]	8.45±0.03	—
AttnGAN[13]	25.83±0.47	35.49
DM-GAN[25]	30.49±0.57	32.64
objGAN[39]	30.29±0.33	—
OP-GAN[40]	28.57±0.17	—
Our CAE-GAN	**30.96±0.56**	**30.83**

表 2 列出了多个模型和本文的 CAE-GAN 模型在 CUB 数据集上的 IS 分数和 FID 分数。表 2 中，在 CUB-200-2011 鸟类数据集上，与经典的 DM-GAN 模型相比，本文设计的 CAE-GAN 模型的 IS 分数从 4.75 左右增长到了 4.87 左右，提升了 2.53%左右，与 2020 年 Gou 等提出的 SegAttnGAN 模型相比，本文模型的 IS 分数也提升了 1.04%左右，说明 CAE-GAN 模型生成的鸟类图片的清晰度有了明显改善。

CAE-GAN 模型在 CUB 数据集上的 FID 分数为 13.66，DM-GAN 的 FID 分数为 16.09，即 CAE-GAN 模型在 DM-GAN 模型的基础上 FID 分数得到明显下降，而 FID 分值越低表示生成图像与真实图像更加接近。因此可以得出，CAE-GAN 模型相比其他方法生成的图片更加接近真实图片，图片内容更加生动形象，图像质量也显著提升。

表 3 列出了 StackGAN 模型[12]、AttnGAN 模型[13]、DM-GAN 模型[25]、objGAN[39]、OP-GAN[40] 和本文的 CAE-GAN 模型在 coco 数据集上的 IS 分数和 FID 分数。表 3 中，在 coco 数据集上，CAE-GAN 模型的 IS 分数为 30.96 左右，DM-GAN 模型的 IS 分数为 30.49 左右，CAE-GAN 模型在 DM-GAN 模型的基础上 IS 分数略有提升，表明 CAE-GAN 模型在 coco 数据集上的生成图片的清晰度和多样性都有了一定提升。

CAE-GAN 模型在 coco 数据集上的 FID 分数为 30.83，DM-GAN 模型在 coco 数据集上的 FID 分数为 32.64，由此可以说明，相比 DM-GAN 模型，CAE-GAN 模型的 FID 分数下降了 5.54%左右，表明本文的 CAE-GAN 模型在 coco 数据集上的生成图像与真实图像更加接近，提升了生成图片的逼真度。

通过对上述实验进行定量分析可以得出，文本的 CAE-GAN 模型生成的图片质量比其他方法生成的图片质量更好，且图片的清晰度、逼真度以及细粒度都有提升，图片的内容也更加逼近真实图像，由此验证了本文设计的 CAE-GAN 网络模型在从文本描述生成图像任务中展现了较好的性能。

为了进一步说明本文的交叉注意力机制对模型性能提升的有效性，本文不采用添加了交叉注意力编码机制的编码器，而采用去掉了交叉注意力编码模块的文本编码器，即预训练仅对文本描述进行直接编码的编码器。然后将其作为 CAE-

GAN 生成对抗网络的预训练编码器进行加载，将该模型命名为 N-CAEGAN，实验结果如表 4、表 5 所列。

表 4 不同模型在 CUB 数据集上的消融实验

Table 4 Ablation experiment of different models on the CUB dataset

Models	IS	FID
StackGAN[12]	3.70±0.04	35.11
AttnGAN[13]	4.36±0.03	23.98
DM-GAN[25]	4.75±0.07	16.09
SegAttnGAN[38]	4.82±0.05	—
N-CAEGAN	4.70±0.05	15.60
Our CAE-GAN	**4.87±0.06**	**13.66**

表 5 不同模型在 coco 数据集上的消融实验

Table 5 Ablation experiment of different models on the coco dataset

Models	IS	FID
StackGAN[12]	8.45±0.03	—
AttnGAN[13]	25.83±0.47	35.49
DM-GAN[25]	30.49±0.57	32.64
objGAN[39]	30.29±0.33	—
OP-GAN[40]	28.57±0.17	—
N-CAEGAN	30.46±0.54	32.58
Our CAE-GAN	**30.96±0.56**	**30.83**

从表 4、表 5 所列的实验结果可知，将去掉交叉注意力编码的文本编码器作为整个文本生成图像网络的预训练编码器加载后，N-CAEGAN 模型在 CUB 数据集和 coco 数据集上的 IS 分数相比本文的 CAE-GAN 模型而言明显下降，而 FID 分数明显升高，说明去除交叉注意力机制在一定程度上会降低生成图片的质量，即充分说明了交叉注意力编码的有效性。

4.4.2 定性评价

为了直观地感受 CAE-GAN 的性能，本文以示例的方式将 CAE-GAN 模型生成的图片同 AttnGAN 模型[13] 和 DM-GAN 模型[25] 生成的图片进行可视化对比，对比结果如图 4、图 5 所示。

图 4 为 CUB 数据集上 3 种不同模型生成的图片示例，可以看出，AttnGAN 模型生成的图片虽然轮廓比较清晰，但细节和纹理比较粗糙，图片的分辨率不高，且背景和实物分界不够明晰，忽略了某些鸟类的丰富细节，如鸟类的脚和羽毛颜色等，生成的图片有些模糊。DM-GAN 模型生成的图片在 AttnGAN 模型的基础上提升了纹理细节，但背景和实物依旧模糊，图片的分辨率有待提高。本文的 CAE-GAN 模型不仅生成了高分辨率的图像，而且图像的轮廓和结构清晰，实物与背景的边界明晰，同时也具备了丰富的细节和细腻的纹理，生成的鸟类的形态也更加具有真实性。

图 5 为 3 种不同模型在 coco 数据集上生成的图片示例，可以看出，AttnGAN 模型生成的图片只有大致的轮廓，图片中场景内容识别困难，生成图片的并没有很好地捕捉到文本描述中提到的图片细节特征，并且图片的分辨率不高，图像比较模糊。DM-GAN 模型生成的图片比 AttnGAN 模型生成的图片轮廓更清晰，能够大致识别场景内容，图片的分辨率较高，但生成图片内容不够完整；生成图片虽然能够捕捉到一些文本描述细节特征，但细节纹理不够突出，细粒度不足，图片的真实性还有待提高。本文的 CAE-GAN 模型生成图像的轮廓清晰，相比前两个模型生成的图片结构更加完整，同时图片的清晰度有了明显的改善，突出了文本描述中提到的细节特征和纹理，质量也得到了显著提升。

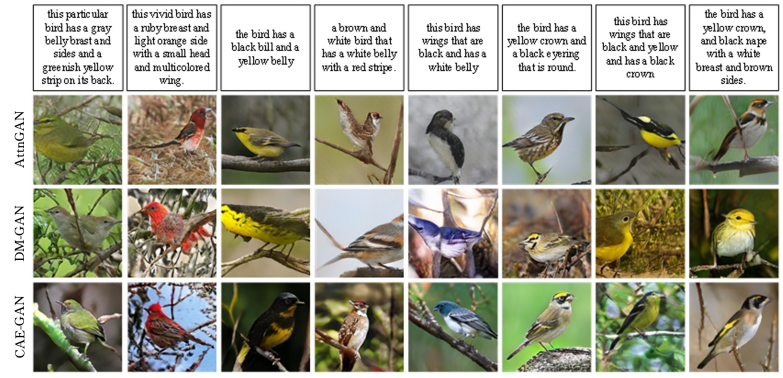

图 4 AttnGAN 模型[13]、DM-GAN 模型[25] 和 CAE-GAN 模型在 CUB 数据集上的生成图像

Fig. 4 Generated images of AttnGAN model, DM-GAN model and CAE-GAN model on CUB dataset

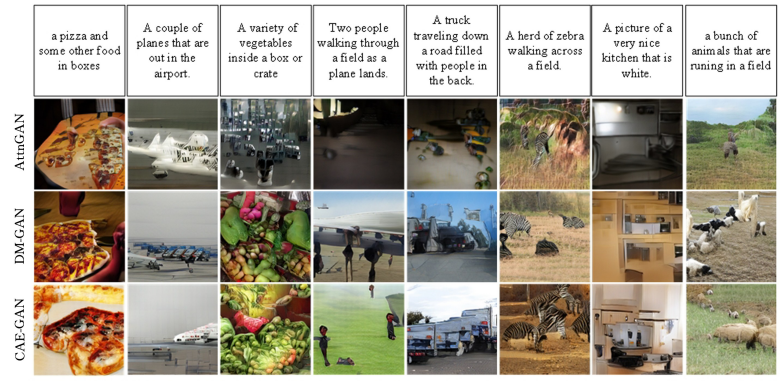

图 5 AttnGAN 模型[13]、DM-GAN 模型[25] 和 CAE-GAN 模型在 coco 数据集上的生成图像

Fig. 5 Generated images of AttnGAN model, DM-GAN model and CAE-GAN model on coco dataset

从图4、图5中可以看出，本文的CAE-GAN模型生成的图片内容基本符合文本描述，与DM-GAN模型和AttnGAN模型相比，CAE-GAN模型生成的图片捕捉到了文本描述中的细节描写特征，生成图片的细节特点更加突出，细节纹理更加丰富，且细粒度有所提升，生成图像的内容更加生动、形象，由此可以看出，CAE-GAN模型生成图像的质量明显高于另外两种模型。

结束语 本文设计了一种以交叉注意力编码为基础的从文本生成图像的方法(CAE-GAN)，通过引入交叉注意力编码器，在预训练编码器阶段对文本信息和图像信息进行交叉联合编码，得到一个全新的交叉注意力编码特征向量，将此向量作为生成对抗网络的输入，经三级生成器逐级生成逼真图像。在预训练编码阶段，传统编码器仅对文本描述进行编码，忽略了文本空间和图像空间之间的语义鸿沟，而本文的交叉联合编码能够捕捉到文本特征和视觉特征之间的内在联系，改善了生成图像的细节，提升了生成图像的质量。由实验结果可知，本文的CAE-GAN模型能够有效地生成高质量的图像。然而，在此基础上如何进一步提升生成图像的多样性仍然是一个有待研究的问题。

参 考 文 献

[1] VINYALS O, TOSHEV A, BENGIO S, et al. Show and tell: A neural image caption generator[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3156-3164.

[2] KARPATHY A, LI F F. Deep visual-semantic alignments for generating image descriptions[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 3128-3137.

[3] ANTOL S, AGRAWAL A, LU J, et al. Vqa: visual question answering[C]// Proceedings of the International Conference on Conputer Vision. 2015: 2425-2433.

[4] ANTOL S, AGRAWAL A, LU J, et al. Vqa: visual question answering[C]// Proceedings of the International Conference on Conputer Vision. 2015: 2425-2433.

[5] JOHNSON J, HARIHARAN B, MAATEN L V D, et al. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2901-2910.

[6] XU K, BA J, KIROS R, et al. Show, attend and tell: Neural image caption generation with visual attention[C]// Proceedings of the 32nd International Conference on International Conference on Machine Learning. Lille, France, 2015: 2048-2057.

[7] WEI Y, ZHAO Y, LU C, et al. Cross-modal retrieval with CNN visual features: A new baseline[J]. IEEE Transactions on Cybernetics, 2016, 47(2): 449-460.

[8] BI J Q, LIU M F, HU H J, et al. Image captioning based on dependency syntax[J]. Journal of Beijing University of Aeronautics and Astronautics, 2021, 47(3): 431-440.

[9] CHEN M J, LIN G J, HAN Q, et al. Asymmetric Patches Nonlocal Total Variation Model for Image Recovery[J]. Journal of Chongqing University of Technology (Natural Science), 2020, 34(2): 127-132, 202.

[10] XU F, MA X P, LIU L B. Cross-modal retrieval method for thyroid ultrasound image and text based on generative adversarial network[J]. Journal of Biomedical Engineering, 2020, 37(4): 641-651.

[11] REED S, AKATA Z, MOHAN S, et al. Learning what and where to draw[OL]. https://arxiv.org/pdf/1610.02454.pdf.

[12] ZHANG H, XU T, LI H S, et al. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks[C]// Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice, Italy, 2017: 5907-7363.

[13] XU T, ZHANG P, HUANG Q, et al. AttnGAN: Fine-Grained text to image generation with attentional generative adversarial networks[C]// Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 1316-1324.

[14] SUN Y, LI L Y, YE Z H, et al. Text-to-image synthesis method based on multi-level structure generative adversarial networks [J]. Journal of Computer Applications, 2019, 39(11): 3204-3209.

[15] XU Y N, HE X H, ZHANG J, et al. Text-to-image synthesis method based on multi-level progressive resolution generative adversarial networks [J]. Journal of Computer Applications, 2020, 40(12): 3612-3617.

[16] MO J W, XU K L, LIN L P, et al. Text-to-image generation combined with mutual information maximization[J]. Journal of Xidian University, 2019, 46(5): 180-188.

[17] HUANG Y W, ZHOU B, TANG X. Text Image Generation Method with Scene Description [J]. Laser & Optoelectronics Progress, 2021, 58(4): 190-198.

[18] WAH C, BRANSON S, WELINDER P, et al. The Caltech-UCSD Birds 200－2011 Dataset[J]. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[19] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft coco: Common objects in context[C]// ECCV. 2014.

[20] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[C]// Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada, 2014: 2672-2680.

[21] MIRZA M, OSINDERO S. Conditional generative adversarial nets[J]. arXiv: 1411.1784, 2014.

[22] NILSBACK M E, ZISSERMAN A. Automated flower classification over a large number of classes[C]// Proceedings of the 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing. Bhubaneshwar, India, 2008: 722-729.

[23] REED S, AKATA Z, YAN X, et al. Generative adversarial text-to-image synthesis[C]// ICML. 2016.

[24] 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,2017,41(8):1947-1962.

[25] 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 the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:5802-5810.

[26] HUANG H Y,GU Z F. A generative adversarial network based on self-attention mechanism for text-to-image generation[J]. Journal of Chongqing University,2020,43(3):55-61.

[27] JU S B,XU J,LI Y F. Text-to-single image method based on self attention[OL]. http://kns. cnki. net/kcms/detail/11. 2127. TP. 20210223. 1347. 018. html.

[28] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need[J]. arXiv:1706. 03762,2017.

[29] LI G,DUAN N,FANG Y J,et al. Unicoder-vl:A universal encoder for vision and language by cross-modal pre-training[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020:11336-11344.

[30] WANG Z H,LIU X H,LI H S,et al. Camp:Cross-modal adaptive message passing for text-image retrieval[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019:5764-5773.

[31] LI L H,YATSKAR M,YIN D,et al. Visualbert:A simple and performant baseline for vision and language[J]. arXiv:1908. 03557,2019.

[32] LU J,BATRA D,PARIKH D,et al. Vilbert:Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks[J]. arXiv:1908. 02265,2019.

[33] TAN H,BANSAL M. Lxmert:Learning cross-modality encoder representations from transformers[J]. arXiv:1908. 07490,2019.

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

[35] SZEGEDY C,ANHOUCKE V V,IOFFE S,et al. Rethinking the inception architecture for computer vision[C]// IEEE. IEEE,2016:2818-2826.

[36] SALIMANS T,GOODFELLOW I J,ZAREMBA W,et al. Improved techniques for training gans[C]// NIPS. 2016.

[37] HEUSEL M,RAMSAUER H,UNTERTHINER T,et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium [C]// NIPS. 2017:6626-6637.

[38] GOU Y C,WU Q C,LI M H,et al. SegAttnGAN:Text to Image Generation with Segmentation Attention [J]. arXiv: 2005. 12444,2020.

[39] LI W,ZHANG P,ZHANG L,et al. Object-driven text-to-image synthesis via adversarial training[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2019:12174-12182.

[40] HINZ T,HEINRICH S,WERMTER S. Semantic object accuracy for generative text-to-image synthesis [J]. arXiv: 1910. 13321,2020.

TAN Xin-yue, born in 1997, postgraduate. Her main research interests include image generation and so on.

HE Xiao-hai, born in 1964, Ph. D, professor, Ph. D supervisor. His main research interests include image processing, pattern recognition and image communication.

(责任编辑:李亚辉)