文章编号：2095-2163(2022)11-0009-09 中图分类号：TP391 文献标志码：A

基于多模态融合的文本生成图像

叶 龙，王正勇，何小海
（四川大学 电子信息学院，成都 610065）

摘 要：生成对抗网络在近几年发展迅速，文本语义和视觉语义之间的联系是文本生成图像的关键，使用生成对抗网络能够生成和文本相匹配的逼真图片。如今，传统的方法是只使用文本编码的方式预训练来对文本进行编码，但是这种算法并没有考虑到与相对应的图像进行语义匹配，而是将输入的文字单独编码，这就导致文本和图像存在语义上的差异性。因此，本文提出了一种基于多模态融合的文本生成图像网络（MLT-GAN）。通过对齐文字信息和视觉信息，来实现图像和文本之间的交互，提高了生成图像的逼真性以及和输入文本的匹配性。实验结果，在 Coco 数据集和 CUB 数据集上，相较于 DM-GAN 模型，本文提出的 MLT-GAN 模型的 *FID* 分数降低了 4.66%和 5.16%，*IS* 指标提高了 1.41%和 1.68%，证明了此方法的有效性。

关键词：生成对抗网络；文本描述；多模态融合；文本生成图像；语义匹配

Text generated images based on multimodal fusion

YE Long, WANG Zhengyong, HE Xiaohai
（College of Electronic and Information Engineering, Sichuan University, Chengdu 610065, China）

【Abstract】The generative adversarial network has developed rapidly in recent years. The relationship between text semantics and visual semantics is the key to text generated images. Using the generative adversarial network can generate realistic images that match the text. Nowadays, the traditional method is to code texts only by pre training in the way of text encoding. However, this algorithm does not consider semantic matching with the corresponding image, but encodes the input texts separately, which leads to semantic differences between texts and images. Therefore, this paper proposes a text generated image network (MLT-GAN) based on multimodal fusion. The interaction between images and texts is realized by aligning text information and visual information, which improves the fidelity of the generated image and the matching with the input text. The experimental results show that compared with DM-GAN model, the MLT-GAN model proposed in this paper reduces the *FID* score by 4.66% and 5.16%, and the *IS* index increases by 1.41% and 1.68% on Coco dataset and CUB dataset. The experimental results prove the effectiveness of this method.

【Key words】Generative Adversarial Network; text description; multimodal fusion; text generated images; semantic matching

0 引 言

文本生成图像[1]属于自然语言处理和计算机视觉的融合任务，是图像生成技术的热点研究课题之一。文本生成图像指从给定的自然语言描述中生成真实的和文本一致的图像。文本生成图像可应用于图像描述生成[2-3]、视觉推理[4]、视觉问答[5]、医疗图像生成[6]等多个领域。

近年来，随着深度学习的快速发展，文本生成图像的主流方法采用生成对抗网络。早期，Mirza 等人[7]提出 CGAN，Reed 等人[8]提出 GAN-INT-CLS，但是使用这些方法生成的图像的质量和分辨率都较低。为了解决生成的图像分辨率的问题，Zhang 等人[9]提出了 Stack-GAN，主要是将生成高分辨率的图像过程分成 2 个阶段。低分辨率的图像是在第一阶段生成，第一阶段主要关注图像的整体结构；第二阶段生成高分辨率的图像，这个阶段主要关注图像的一些细节信息以及纠正第一阶段生成图像的一些错误。

多阶段图像生成的方法虽然解决了生成图像分辨率低的问题，但是生成的图像和输入文本依然存在语义匹配较低的问题。AttnGAN[10]引入注意力机制，通过注意力把生成图像和句子特征向量中最密切的部分联系起来。DM-GAN[11]通过引入动态记忆化机制来使得初始图像自适应地选择重要的文本信息，但是依然存在生成图像缺失、生成图像质量不高、低分辨率阶段生成图像与文本描述不相符的问题。

针对上述问题，本文提出了一种基于多模态融合的文本生成图像方法，在图像特征提取和文本描

作者简介：叶 龙(1995-)，男，硕士研究生，主要研究方向：计算机视觉；王正勇(1969-)，女，博士，副教授，硕士生导师，主要研究方向：图像处理与模式识别、通信与信息处理、计算机视觉；何小海(1964-)，男，博士，教授，博士生导师，主要研究方向：图像处理、模式识别、图像通信。
通讯作者：王正勇 Email：wangzhengy@scu.edu.cn

收稿日期：2022-07-05

述提取时采用通道注意力来突出重要信息,同时将提取出的文本特征和图像特征用双线性池化[12]进行融合,从而得到文本信息和对应图像信息之间的映射关系。

1 相关工作

1.1 通道注意力机制

近年来,通道注意力在视觉处理[13]等任务得到广泛应用,其基本原理是通过对每个特征通道进行加权,来突出关键信息、抑制无效信息,从而达到提高特征表示能力的目的。Hu 等人提出了 SENet[14],SENet 使用全局损失函数来自适应地调整每个通道的权重,SENet 在图像分类方面效果显著。

1.2 多模态融合注意力机制

AttnGAN 中加入了注意力来提升文本生成图像的质量,但是,文本信息和图像信息之间的交互对于文本生成图像是至关重要的,特别是文本特征和图像特征之间的联系以及对齐。最近,双线性池化(MFB)在视觉问答方面表现出很好的效果,视觉问答需要做的是同时理解图像内容和文本内容,文本生成图像同样也需要理解图像内容和文本内容,因此,采用 MFB 将文本信息和图像信息进行融合编码,这种多模态融合编码能够有效提升生成图像的质量。

1.3 文本生成图像方法

文本生成图像主流的方法是使用堆叠式网络来生成高质量的图像。Zhang 等人[9]提出了 StackGAN,采用了 2 个堆叠的生成器,第一阶段关注图像的背景、轮廓等基本信息,生成低分辨率的 64 * 64 像素的图片,第二阶段弥补之前缺失的细节和纹理等高级特征,生成 256 * 256 高分辨率的图像。Xu 等人[10]提出了 AttnGAN 模型,该模型在生成网络中引入了自注意力机制,AttnGAN 实现了单词与图片中的某个子区域的对应,自动选择字级条件以生成图像不同子区域。2019 年,Qiao 等人[15]提出了 MirrorGAN 来实现图像到文本,文本到图像的双重映射。Zhu 等人[11]提出的 DMGAN 通过引入动态记忆化机制来使得初始图像自适应地选择重要的文本信息。然而现有的对文本编码的方式,没有考虑到文本信息与对应图像之间的映射关系,导致第一阶段生成的图像和输入文本的不匹配,也会导致后面两级图像的优化受到影响。因此,本文基于 DM-GAN 网络进行改进,在图像特征提取和文本描述提取时采用通道注意力来突出重要信息,在预训练文本编码器时引入了双线性池化,将文本特征和图像特征进行联合编码后,输出一个新的融合后的特征向量,新的特征向量学习到图像和文本之间的关系,因此可以生成更加真实的图像。

2 基于多模态融合的生成对抗网络

MLT-GAN 模型框架如图 1 所示。由图 1 可知,本文设计的 MLT-GAN 由预训练编码器、生成对抗网络和动态存储三个模块构成。多模态融合注意力机制用于预训练编码器,是将文本特征输入到多模态融合编码器中,多模态融合编码器将输出特征向量 $\boldsymbol{f}_e$ 和单词特征矩阵 $\boldsymbol{W}$。随机噪声和多模态融合注意力向量相结合,输入到生成对抗网络中,三级生成器逐级生成高分辨率的图像。单词特征矩阵 $\boldsymbol{W}$ 主要是用来在动态存储模块中和初级图像特征进行融合来生成下一级的图像特征。上述过程的数学方法公式分别如下:

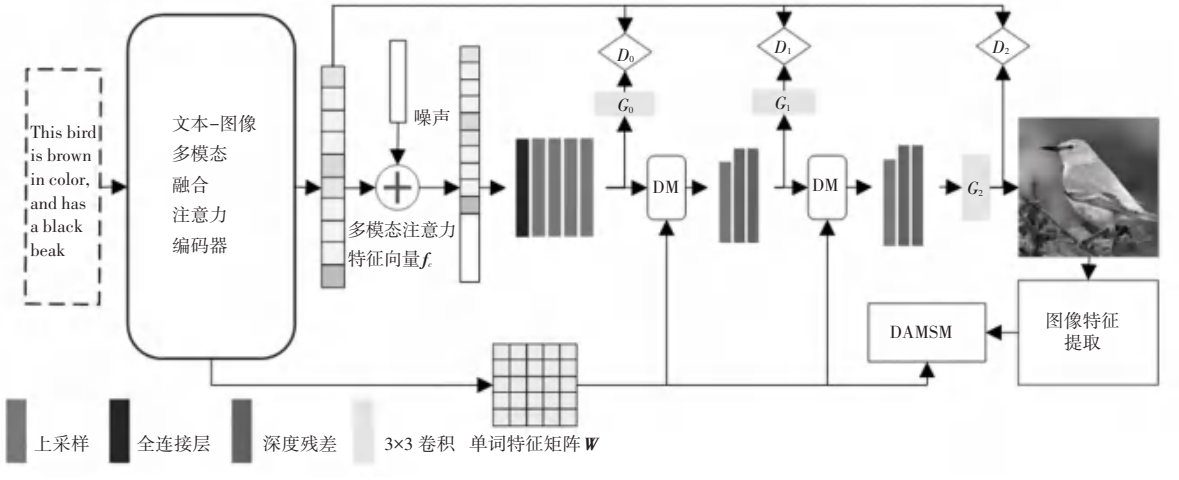

图 1 MLT-GAN 模型框架图

Fig. 1 MLT-GAN model framework diagram

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

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

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

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

其中，C_E 是多模态融合编码器；DM 是动态存储模块；原始图像特征是 F_R；G_0,G_1,G_2 表示三级生成器；$\boldsymbol{s}$ 是从文本描述中提取的全局句子向量；$\boldsymbol{F}_0$，$\boldsymbol{F}_1,\boldsymbol{F}_2$ 是 G_0,G_1,G_2 生成的图像特征；$\boldsymbol{z}$ 是随机高斯噪声。

2.1　多模态融合注意力编码

本文设计了一种多模态融合编码器来将图像信息和文本信息进行联合编码和对齐。

多模态融合编码器框架如图 2 所示。由图 2 可看到，多模态融合编码器由 4 部分组成，包括文本特征提取、图像特征提取、通道注意力编码和多模态融合注意力编码。对此拟展开研究分述如下。

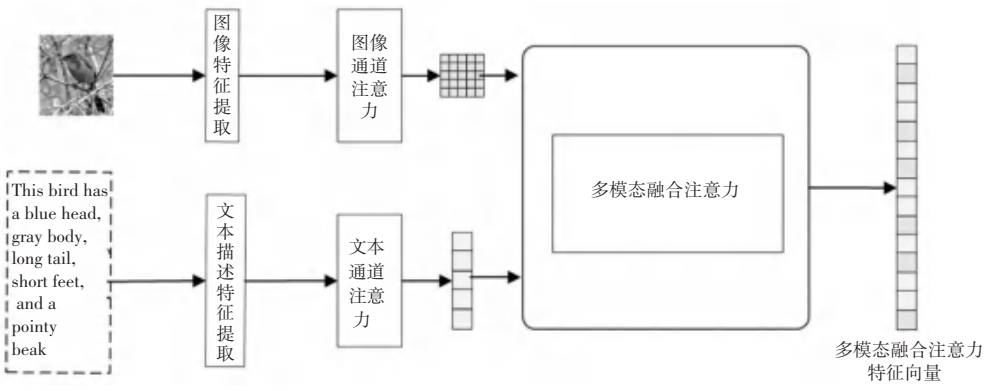

图 2　多模态融合编码器框架图

Fig. 2　Multimodal fusion encoder framework diagram

（1）文本特征提取。提取文本特征用的是双向长短时网络[12]（LSTM），双向长短时网络是将文本描述进行编码，输出一个单词特征矩阵 $\boldsymbol{W}^{d*t}$ 和全局句子特征向量 $\boldsymbol{s}$。推得的数学公式为：

$$\boldsymbol{W}^{d*t},\ \boldsymbol{s} = T_E(\boldsymbol{T}_{text}) \tag{5}$$

其中，t 表示单词的个数；d 表示词向量的维度；$\boldsymbol{T}_{text}$ 表示原文本描述；T_E 表示双向 LSTM 网络。

（2）图像特征提取。图像特征提取采用 InceptionV3 模型[13]。此处需用到的公式为：

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

（3）通道注意力编码。为了突出图像特征和文本描述特征中的重要信息，引入通道注意力，将特征提取后的图像特征图和文本特征向量输入到通道注意力中，采用通道注意力对图像特征图和文本特征向量进行加权，使得生成的图像多样性更加丰富。图像通道注意力和文本通道注意力如图 3、图 4 所示。

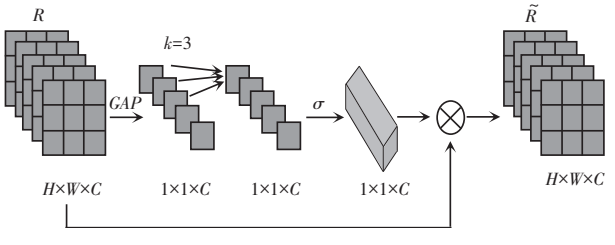

图 3　图像通道注意力模块

Fig. 3　Image channel attention module

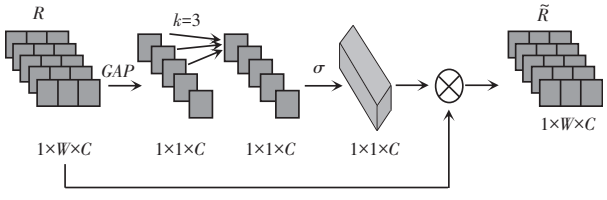

图 4　文本通道注意力模块

Fig. 4　Text channel attention module

在通道注意力模块中，权重 $\boldsymbol{w}$ 的计算如下：

$$\boldsymbol{w} = \sigma(\boldsymbol{Q}y) \tag{7}$$

其中，$y = G_{GAP}(R)$，是通过对输入的特征图经过平均池化后得到；σ 是 *Sigmoid* 函数；$\boldsymbol{Q}$ 是权重矩阵。

假定接受的特征图 $R \in R^{W\times H\times C}$，$W$、$H$、$C$ 分别表示特征图的宽度、高度和通道维度。全局平均池化的计算公式如下：

$$G_{GAP}(R) = \frac{1}{WH}\sum_{i=1,j=1}^{W,H} R_{i,j} \tag{8}$$

权重矩阵 $\boldsymbol{Q}$ 的尺寸是 $k \times C$，针对每一个通道 y_i，对应的权重 w_i，仅需考虑相邻的 k 个通道的相应加权（本文设置的是 3），如下式所示：

$$w_i = \sigma\Big(\sum_{j=1}^{k} w^j y_i^j\Big) \quad y_i^j \in \Omega_i^k \tag{9}$$

其中，Ω_i^k 表示通道 y_i 相邻接的 k 个通道的集合，w^j 表示 y_i 的第 j 个相邻通道 y_i^j 的权重。这里给出的数学计算公式具体如下：

$$s' = s \times w_i \tag{10}$$

$$f_v' = f_v \times w_i \tag{11}$$

其中，s' 和 f_v' 分别是通过通道注意力的全局句子特征向量和图像特征。

(4)多模态融合注意力编码。多模态融合注意力编码主要是将文本特征和图像特征的内部联系搭建起来，实现两者的联合编码。经过通道注意力的图像特征 f_v' 和全局句子特征 s' 通过多模态融合注意力编码后，融合成一个新的特征 f_c，本文采用的多模态融合注意力编码方法是双线性池化(Bilinear Pooling)。数学函数形式见如下：

$$f_c = Bilinear\ Pooling(s', f_v') \tag{12}$$

双线性池化具体细节如图 5 所示。由图 5 可看到，双线性池化可以分解为 2 个阶段，首先，不同模态的特征被扩展到高维空间，然后进行元素相乘，接着经过总和池化获取向量的全局特征，再通过归一化层来将高维特征进行压缩输出。

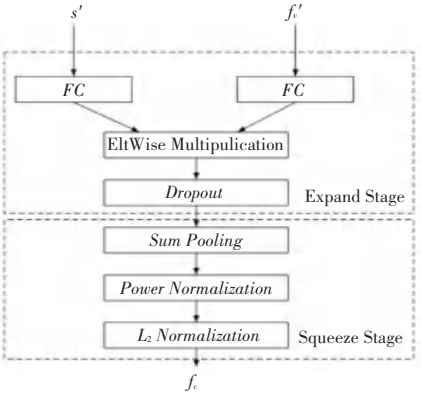

图 5 双线性池化

Fig. 5 Bilinear pooling

2.2 经典三级生成对抗网络

由图 1 可知，MLT-GAN 模型采用了和 StackGAN、StackGAN++、AttnGAN、DM-GAN 相类似的三级对抗生成网络，分别为 G_0/D_0，G_1/D_1，G_2/D_2。G_0 由一个大小为 3 * 3 的卷积层、3 个上采样层和一个全连接层组成，第一阶段生成 64×64 分辨率的图像；第二阶段 G_1 和 G_2 在 G_0 的基础上进行优化，分别生成 128 * 128 分辨率的图像和 256 * 256 分辨率的图像，两者的结构一致，由 2 个深度残差网络层、1 个上采样层和 1 个大小为 3 * 3 的卷积网络层组成。

2.3 动态存储记忆模块

动态存储模块存在于生成器 G_0 与 G_1，生成器 G_1 与生成器 G_2 之间，该模块的作用是在初始图像的生成上，基于动态内存将图像质量进行进一步的细化。动态存储模块框图如图 6 所示。图 6 中，动态存储记忆模块由 4 部分组成，分别为：内存写入、键寻址、值读取、响应。研究对此将给出探讨论述如下。

(1)模块的输入是：

$$\boldsymbol{W} = \{w_1, w_2, \ldots, w_T\},\ w_i \in R^{N_w} \tag{13}$$

$$R_i = \{r_1, r_2, \ldots, r_N\},\ r_i \in R^{N_r} \tag{14}$$

其中，$\boldsymbol{W}$ 表示单词特征矩阵；R_i 表示图像特征；R_0 表示初始图像特征；R_1 表示第二级图像特征；R_2 表示第三级图像特征；T 表示单词个数；N_w 表示单词特征维数；N 表示图像像素个数；N_r 表示图像像素特征矩阵的维度。

(2)内存写入门。主要通过内存写门来实现，通过选择相关单词来细化初始化图像，对此可表示为：

$$g_i^w(R, w_i) = \sigma(\boldsymbol{A} * w_i + \boldsymbol{B} * \frac{1}{N}\sum_{i=1}^{N} r_i) \tag{15}$$

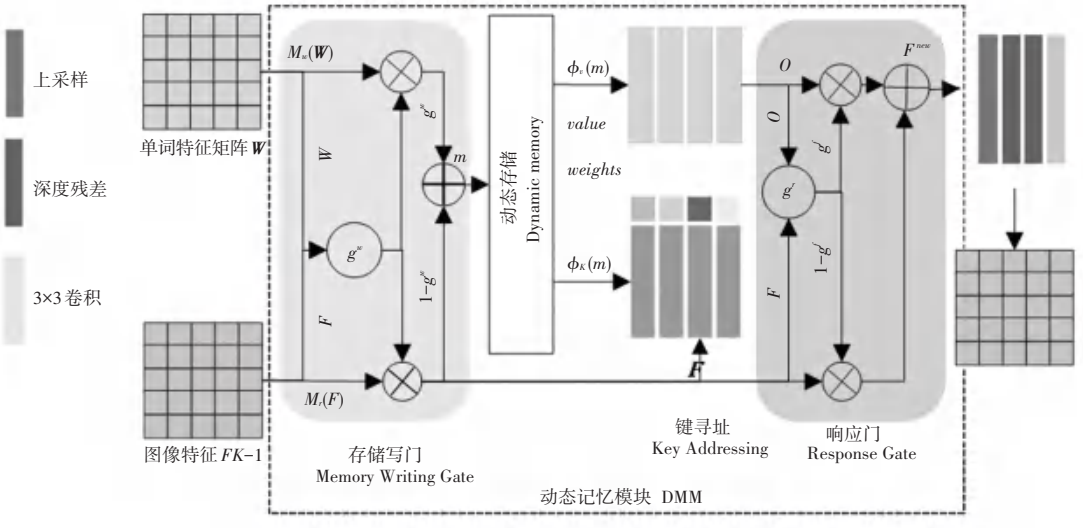

图 6 动态存储模块框图

Fig. 6 Dynamic storage block diagram

$$m_i = M_w(w_i) * g_i^w + M_r(\frac{1}{N}\sum_{i=1}^{N} r_i) * (1 - g_i^w) \quad (16)$$

其中，σ 表示 *sigmoid* 函数；$\boldsymbol{A}$ 为 $1 * N_w$ 维矩阵；$\boldsymbol{B}$ 为 $1 * N_r$ 维矩阵；$M_w(\cdot)$ 和 $M_r(\cdot)$ 表示 $1 * 1$ 的卷积操作，$M_w(\cdot)$ 和 $M_r(\cdot)$ 是以 N_m 维度把文字特征和图像特征嵌入到同一个特征空间中。

(3)键寻址过程。在这一步中，使用密钥存储器检索相关的存储器，计算每个内存槽的权重，作为内存槽 m_i 与图像特征 r_j 的相似概率，可由如下公式来求值：

$$a_{i,j} = \frac{\exp(\phi_K(\boldsymbol{m}_i)^{\mathrm{T}}\boldsymbol{r}_j)}{\sum_{l=1}^{T}\exp(\phi_K(\boldsymbol{m}_l)^{\mathrm{T}}\boldsymbol{r}_j)} \quad (17)$$

其中，$a_{i,j}$ 表示第 i 个内存和第 j 个图像特征的相似度，$\phi_K()$ 是 $1 * 1$ 的卷积网络，目的是将内存特征映射到 N_r 维度。

(4)值读取过程。输出记忆表示定义为根据相似概率的记忆加权求和，数学定义公式具体如下：

$$o_j = \sum_{i=1}^{T} a_{i,j}\phi_V(m_i) \quad (18)$$

其中，$\phi_V()$ 为值内存访问进程，将内存特性映射到 N_r 维数，$\phi_V()$ 实现 $1 * 1$ 的卷积操作。

(5)响应门。是用来完成响应步骤的，响应门是通过利用门控机制来及时控制信息以及图像信息的更新。可由如下公式进行描述：

$$g_i^r = \sigma(\boldsymbol{W}[o_i, r_i] + \boldsymbol{b}) \quad (19)$$

$$r_i^{new} = o_i * g_i^r + r_i * (1 - g_i^r) \quad (20)$$

其中，g_i^r 为信息融合的响应门；σ 为 *sigmoid* 函数；$\boldsymbol{W}$ 是参数矩阵；$\boldsymbol{b}$ 是偏置项。

2.4　损失函数

MLT-GAN 的损失函数由 2 部分组成，分别为生成器损失函数和判别器损失函数。文中对此可做阐释解析如下。

(1)生成器损失函数 L。由 3 部分组成：分别为条件损失函数 L_{CA}、生成损失函数 L_{G_i} 和深度多模态相似模型损失函数(DAMSM) L_{DAMSM}。即可由下式来计算：

$$L = \sum_i L_{G_i} + \lambda_1 L_{CA} + \lambda_2 L_{DAMSM} \quad (21)$$

其中，λ_1 和 λ_2 分别为条件损失 L_{CA} 和深度多模态相似模型损失函数 L_{DAMSM} 的权重。

① L_{G_i} 和 L_{CA}。推导得到的公式分别为：

$$L_{G_i} = -\frac{1}{2}[E_{x\sim pG_i}\log D_i(x) + E_{x\sim pG_i}\log D_i(x,s)] \quad (22)$$

$$L_{CA} = D_{KL}(N(u(\boldsymbol{s}), \sum(\boldsymbol{s})) \| N(0,I)) \quad (23)$$

其中，$u(s)$ 是句子特征的均值，$\sum(\boldsymbol{s})$ 是句子对角协方差矩阵。$u(\boldsymbol{s})$ 和 $\sum(\boldsymbol{s})$ 由全连接层计算，式(22)中，第一项是无条件损失，目的是使得生成的图像尽可能真实，第二项是条件损失，目的是使得图像与输入的句子相符合。条件损失 L_{CA} 用来防止过拟合。

② L_{DAMSM}。DAMSM 损失函数用来衡量图像和文本描述的匹配程度，DAMSM 损失函数使生成的图像更好地适应文本描述。

(2)判别器损失函数。由条件损失 L_{CD} 和非条件损失 L_D 组成，具体公式如下：

$$L_{D_i} = -\frac{1}{2}[L_D + L_{CD}] \quad (24)$$

其中，

$$L_D = E_{x\sim Pdata}\log D_i(x) + E_{x\sim PG_i}\log(1 - D_i(x)) \quad (25)$$

$$L_{CD} = E_{x\sim Pdata}\log D_i(x,s) + E_{x\sim PG_i}\log(1 - D_i(x,s)) \quad (26)$$

其中，无条件损失 L_D 是用来区分生成的图像和真实图像，条件损失 L_{CD} 是用来判断输入的句子和图像是否符合。

3　实验

3.1　实验数据集

本文在 Coco[16] 和 CUB[17] 两个数据集上分别进行了训练和测试。其中，CUB 数据集是专门针对鸟类图像的数据集，CUB 数据集收录了 200 种鸟类，数据集包括鸟类图片和对应的文本描述。Coco 数据集包含了复杂场景、丰富的类别，共有 80 个类别，数据集的具体情况见表 1。

表 1　数据集

Tab. 1　The experimental dataset

数据集	训练集	测试集
CUB	8 855	2 933
Coco	82 783	40 470

3.2　实验过程

本文在公开数据集 Coco 和 CUB 数据集上训练和测试了 MLT-GAN。

实验共由 3 步组成,第一步预训练多模态融合编码器,第二步训练整个模型,第三步测试整体模型的性能效果。对此内容可做重点论述如下。

(1)预测训练多模态融合编码器。通过不同的任务预训练多模态融合编码器,来得到每个任务中文本信息与图像信息之间的关系,可以得到对应此任务的文本与对应的图像的融合编码,运行的结果是保存训练好的编码器模型。

(2)训练整个模型。在整个模型训练过程中,首先加载已经过训练并保存了的编码器模型,接着单独训练 MLT-GAN 模型的剩余部分。

(3)测试整个模型的性能效果。分别在 Coco 数据集和 CUB 数据集上进行测试,本文的 MLT-GAN 均生成了30 000张逼真图像,通过计算相应的 *IS* 分数和 *FID* 分数,来衡量本文提出的 MLT-GAN 模型的性能好坏。

3.3 评价指标

本文采用 *FID*[18] (Frechet Inception Distance)和 *IS*[19](Inception Score)分数来衡量 MLT-GAN 的性能。对此,文中将进行研究表述见如下。

(1) *IS*。*IS* 值越高,表示生成图片的多样性和品质就越好, *IS* 的公式如下:

$$IS = e^{E_{x\sim pG}D_{KL}(p(y|x)\,\|\,p(y))} \tag{27}$$

其中, $p(y\mid x)$ 是预训练图像编码器预测的对应标签 y 的条件概率,$p(y)$ 则是预训练图像编码器预测的对应标签 y 的边缘概率。

(2) *FID* 得分。是指真实图像与虚假图像之间在特征方面的距离,当真实图像与虚假图像特征越近时,*FID* 值就越小。其计算方法为:

$$FID = \|u_r - u_g\|^2 + tr\Big(\sum_r + \sum_g - 2\Big(\sum_r \sum_g\Big)^{\frac{1}{2}}\Big) \tag{28}$$

其中, u_r 是真实图像的均值; $\sum_r$ 是真实图像的特征协方差; u_g 是生成图像的均值; $\sum_g$ 是生成图像的特征协方差。

3.4 实验结果

3.4.1 定量评价

本文从定量评价和定性评价两个方面来评估 MLT-GAN 模型的性能。本文使用在 Coco 数据集和 CUB 数据集的测试集中生成的 30 000 张图像来计算 *FID* 分数和 *IS* 分数,并与一些主流的对抗生成网络进行了对比,实验结果见表 2、表 3。

表 2 不同模型在 CUB 数据集上的 *FID* 和 *IS* 分数

Tab. 2 *FID* and *IS* scores of different models on the CUB dataset

Models	*IS* ↑	*FID* ↓
StackGAN	3.70(±0.04)	35.11
AttnGAN	4.36(±0.03)	23.98
DM-GAN	4.75(±0.07)	16.09
SegAttnGAN[20]	4.82(±0.05)	-
MA-GAN[21]	4.76(±0.09)	21.66
The proposed MLT-GAN	4.83(±0.07)	15.26

表 3 不同模型在 Coco 数据集上的 *FID* 和 *IS* 分数

Tab. 3 *FID* and *IS* scores of different models on the Coco dataset

Models	*IS* ↑	*FID* ↓
AttnGAN	25.83(±0.47)	35.49
DM-GAN	30.49(±0.57)	32.64
ObjGAN[22]	30.29(±0.33)	-
OP-GAN[23]	28.57(±0.17)	-
The proposed MLT-GAN	30.92(±0.32)	31.12

表 2 列出了 MLT-GAN 与部分主流的对抗生成网络在 CUB 数据集上的 *FID* 和 *IS* 分数。与本文的基础网络 DM-GAN 模型相比,本文设计的 MLT-GAN 网络的 *IS* 分数从 4.75 提高到 4.83,可知提升了2.11%,DM-GAN 模型的 *FID* 分数为 16.09,而本文提出的 MLT-GAN 模型的分数为 15.26,显然有所下降,说明本文提出的 MLT-GAN 模型生成的鸟类图像在图像质量和清晰度上有了明显的改善。

表 3 列出了 MLT-GAN 与部分主流的对抗生成网络在 CUB 数据集上的 *FID* 和 *IS* 分数。与本文的基础网络 DM-GAN 模型相比,本文设计的 MLT-GAN 网络的 *IS* 分数从 30.49 提高到 30.92,DM-GAN 模型的 *FID* 分数为 32.64,而本文提出的 MLT-GAN 模型的分数为 31.12,已出现明显的下降,说明本文提出的 MLT-GAN 模型生成的鸟类图像在图像质量和多样性上有了一定的改善。

通过上述实验的定量的分析可得,本文提出的 MLT-GAN 模型所生成的图像质量和清晰度比其他方法生成的图像质量和图像清晰度有了一定的提升,生成图像的内容也更加接近真实的图像,证明了本文提出的 MLT-GAN 模型在文本生成图像任务中具有良好的效果。

为了进一步检验本文所述的通道注意力机制和多模态融合注意力机制在提高模型性能方面的作用,本文将基础网络 DM-GAN 上加入通道注意力模

块，将其命名为 TDM-GAN，将基础网络 DM-GAN 上加入多模态融合注意力模块，将其命名为 MDM-GAN，将本文提出的 MLT-GAN 同其进行对比，实验结果见表 4、表 5。

表 4　不同模型在 CUB 数据集上的消融实验

Tab. 4　Ablation experiments of different models on CUB datasets

Models	*IS* ↑	*FID* ↓
DM-GAN	4.75(±0.07)	16.09
TDM-GAN	4.77(±0.08)	16.02
MDM-GAN	4.80(±0.06)	15.94
The proposed MLT-GAN	4.83(±0.07)	15.26

表 5　不同模型在 Coco 数据集上的消融实验

Tab. 5　Ablation experiments of different models on Coco datasets

Models	*IS* ↑	*FID* ↓
DM-GAN	30.49(±0.57)	32.64
TDM-GAN	30.53(±0.62)	32.26
MDM-GAN	30.68(±0.67)	31.72
The proposed MLT-GAN	30.92(±0.32)	31.12

根据表 4、表 5 给出的实验结果可以得到，本文提出的 MLT-GAN 比去除了通道注意力和多模态融合注意力模块的网络效果更好。

3.4.2　定性评价

为了更加直观评价 MLT-GAN 的性能，本文以示例的形式将 MLT-GAN 模型生成的图像和 AttnGAN 网络模型、DM-GAN 网络模型生成的图像进行对比，对比结果如图 7、图 8 所示。

图 7 是 CUB 数据集上 3 种模型生成的部分图像。从图 7 中可以看出，AttnGAN 和 DM-GAN 生成的图像实物和背景的边界不清晰，存在模糊区域，忽略了鸟类图像的一些细节特征，图像的分辨率不高，而本文提出的 MLT-GAN 生成的鸟类图像背景与实物背景分明，生成的图像分辨率高且具有更多的细节特征。

图 8 是 3 种模型在 Coco 数据集上生成的部分图像。从图 8 中可以看出，AttnGAN 模型生成的图像轮廓不完整，图片中具体的场景很难识别，DM-GAN 模型生成的图像质量相较于 AttnGAN 有了一定的提升，但是生成的图像内容残缺，捕捉到的细节特征不够明显，图片的质量有待提高。而本文提出的 MLT-GAN 模型生成的图像存在较少失真，图像内容结构完整，轮廓清晰，文本描述中的细节和纹理的重点得以突出，图像质量得到显著提高。

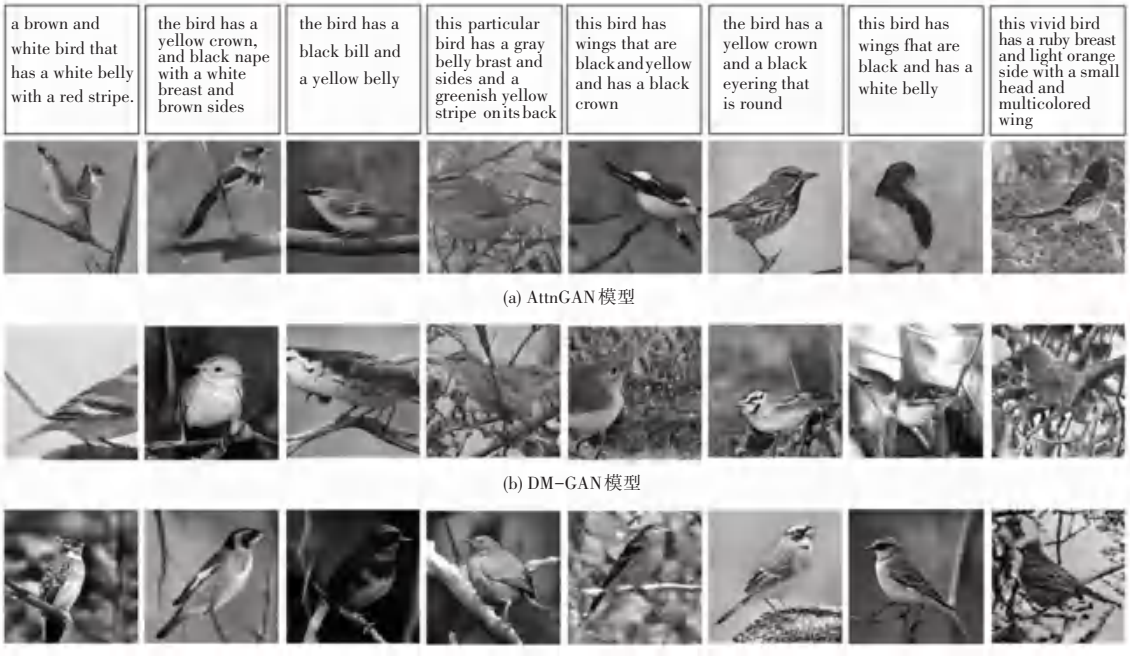

图 7　AttnGAN、DM-GAN、MLT-GAN 在 CUB 数据集上生成的图像

Fig. 7　Generated images of the AttnGAN model、DM-GAN model and MLT-GAN model on the CUB dataset

图 8 AttnGAN、DM-GAN、MLT-GAN 在 Coco 数据集上生成的图像

Fig. 8 Generated images of the AttnGAN model、DM-GAN model and MLT-GAN model on the Coco dataset

4 结束语

本文提出了一种基于多模态融合的文本生成图像方法(MLT-GAN),通过在预训练编码阶段引入通道注意力和多模态融合注意力来对文本信息和图像信息进行融合编码,从而捕捉到文本特征和视觉特征之间的内在联系,提升了图像的质量。实验结果表明,在 Coco 数据集和 CUB 数据集上,相较于 DM-GAN 模型,本文提出的 MLT-GAN 模型的 *FID* 分数降低了 4.66%和 5.16%, *IS* 指标提高了 1.41%和 1.68%。本文提出的 MLT-GAN 在 CUB 数据集和 Coco 数据集,相较于基础网络 DM-GAN 和单独添加了通道注意力的 TDM-GAN 以及单独添加了多模态融合注意力的 MDM-GAN 都有一定的提高,因此,本文提出的 MLT-GAN 在文本生成图像任务中具有良好的效果,生成图片的质量得到显著提高。

参考文献

[1] RUAN Shulan, ZHANG Yong, ZHANG Kun, et al. Dae-gan: Dynamic aspect-aware gan for text-to-image synthesis[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision.Montreal, Canada: IEEE,2021: 13960-13969.

[2] GAO Junlong, WANG Shiqi, WANG Shanshe, et al. Self-critical n-step training for image captioning[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach,USA ,IEEE,2019: 6300-6308.

[3] DOGNIN P, MELNYK I, MROUEH Y, et al. Adversarial semantic alignment for improved image captions[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach,USA ,IEEE,2019: 10463-10471.

[4] GE Yunhao, XIAO Yao, XU Zhi, et al. A peek into the reasoning of neural networks: Interpreting with structural visual concepts [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Kuala Lumpur: IEEE, 2021: 2195-2204.

[5] UROOJ A, KUEHNE H, DUARTE K, et al. Found a reason for me? weakly-supervised grounded visual question answering using capsules[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Kuala Lumpur:IEEE, 2021: 8465-8474.

[6] YI Xin, WALIA E, BABYN P. Generative adversarial network in medical imaging: A review[J]. Medical image analysis, 2019, 58: 101552.

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

[8] REED S, AKATA Z, YAN Xinchen, et al. Generative adversarial text to image synthesis[C]//International conference on machine learning. New York, USA:PMLR, 2016: 1060-1069.

[9] ZHANG Han, XU Tao, LI Hongsheng, et al. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks[C]//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy :IEEE,2017: 5907-5915.

[10] XU Tao, ZHANG Pengchuan, HUANG Qiuyuan, 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.Salt Lake City,USA:IEEE, 2018: 1316-1324.

[11] ZHU Minfeng, PAN Pingbo, CHEN Wei, 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. Long Beach, CA：IEEE, 2019：5802－5810.
[12] YU Zhou, YU Jun, FAN Jianping, et al. Multi－modal factorized bilinear pooling with co－attention learning for visual question answering[C]//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy:IEEE,2017：1821－1830.
[13] WANG Gu, MANHARDT F, TOMBARI F, et al. GDR－net：Geometry－guided direct regression network for monocular 6D object pose estimation [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Kuala Lumpur:IEEE,2021：16611－16621.
[14] SALLOUM R, REN Yuzhuo, KUO C C J. Image splicing localization using a multi－task fully convolutional network (MFCN) [J]. Journal of Visual Communication and Image Representation, 2018, 51：201－209.
[15] QIAO Tingting, ZHANG Jing, XU Duanqing, et al. Mirrorgan：Learning text－to－image generation by redescription [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.Long Beach,USA:IEEE, 2019：1505－1514.
[16] DÖHNER H, ESTEY E H, AMADORI S, et al. Diagnosis and management of acute myeloid leukemia in adults：recommendations from an international expert panel, on behalf of the European LeukemiaNet[J]. Blood, 2010, 115(3)：453－474.
[17] WAH C, BRANSON S, WELINDER P, et al. The caltech－ucsd birds－200－2011 dataset [R]. USA：California Institute of Technology,2011.
[18] 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′17：Proceedings of the 31st International Conference on Neural Information Processing Systems. NY, USA：NIPS Foundation, 2017:6629－6640.
[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. Barcelona：Curran Associates Inc., 2016：2234－2242.
[20] GOU Yuchuan, WU Qiancheng, LI Minghao, et al. SegattnGAN：Text to image generation with segmentation attention [J]. arXiv preprint arXiv:2005.12444, 2020.
[21] YANG Yanhua, WANG Lei, XIE De, et al. Multi－sentence auxiliary adversarial networks for fine－grained text－to－image synthesis[J]. IEEE Transactions on Image Processing, 2021, 30：2798－2809.
[22] LI Wenbo, ZHANG Pengchuan, ZHANG Lei, et al. Object－driven text－to－image synthesis via adversarial training [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA：IEEE, 2019：12174－12182.
[23] HINZ T, HEINRICH S, WERMTER S. Semantic object accuracy for generative text-to-image synthesis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(3)：1552－1565.

(上接第 8 页)

[11] LIU Jixin, XIA Yinyun, TANG Zheng. Privacy-preserving video fall detection using visual shielding information[J]. The Visual Computer, 2021, 37：359－370.
[12] LIU Jixin, TAN Rong, HAN Guang, et al. Privacy－preserving in－home fall detection using visual shielding sensing and private information embedding[J]. IEEE Transactions on Multimedia, 2021, 23：3684－3699.
[13] 王春峰, 李军. 基于面部检测和深度神经网络的面部情绪自动识别算法[J]. 光电子·激光, 2020, 31(11)：1197－1203.
[14] LIU Xiao, CHENG Xiangyi, LEE K. GA－SVM－based facial emotion recognition using facial geometric features[J]. IEEE Sensors, 2021, 21(10)：11532－11542.
[15] SAXENA S, TRIPATHI S, SUDARSHAN T S B. An intelligent facial expression recognition system with emotion intensity classification[J]. Cognitive Systems Research, 2022, 74：39－52.
[16] YANG Yi, GAO Qiang, SONG Yu, et al. Investigating of deaf emotion cognition pattern by EEG and facial expression combination [J]. IEEE Journal of Biomedical and Health Informatics, 2022, 26(2)：589－599.
[17] VIOLA P, JONES M. Rapid object detection using a boosted cascade of simple features[C]// Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). Kauai, HI, USA：IEEE, 2001：8072.
[18] CAI Zhaowei, VASCONCELOS N. Cascade R－CNN：High quality object detection and instance segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(5)：1483－1498.
[19] OJALA T, PIETIKäINEN M, MäENPää T. Multiresolution gray－scale and rotation invariant texture classification with local binary patterns[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(7)：971－987.
[20] WRIGHT J, YANG A Y, GANESH A, et al. Robust face recognition via sparse representation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(2)：210－227.
[21] LIU Jixin, SUN Quansen. Sparse recognition via intra－class dictionary learning using visual saliency information [J]. Neurocomputing, 2016, 196：70－81.
[22] 刘佶鑫, 魏嫚. 可见光-近红外 HSV 图像融合的场景类字典稀疏识别方法[J]. 计算机应用, 2018, 38(12)：3355－3359, 3366.
[23] LIU Jixin, HAN Guang, SUN Ning, et al. Generalized compressed sensing with QR-based vision matrix learning for face recognition under natural scenes[J]. Signal Processing：Image Communication, 2019, 77：11－19.
[24] KAMATH A, BISWAS A, BALASUBRAMANIAN V. A crowdsourced approach to student engagement recognition in e－learning environments [C]//2016 IEEE Conference on Applications of Computer Vision. Piscataway：IEEE, 2016：1－9.
[25] SHEIKH H R, SABIR M F, BOVIK A C. A statistical evaluation of recent full reference image quality assessment algorithms[J]. IEEE Transactions on Image Processing, 2006, 15(11)：3440－3451.
[26] HUANG G B, MATTAR M, BERG T, et al. Labeled faces in the wild：A database forstudying face recognition in unconstrained environments[R]. TAmherst：University of Massachusetts, 2007.