DOI：10.11830/ISSN.1000-5013.201807038

分类重构堆栈生成对抗网络的文本生成图像模型

陈鑫晶，陈锻生

（华侨大学 计算机科学与技术学院，福建 厦门 361021）

摘要： 利用堆栈生成对抗网络，提出分类重构堆栈生成对抗网络. 第一阶段生成 64 px×64 px 的图像，第二阶段生成 256 px×256 px 的图像. 在每个阶段的文本生成图像中，加入图像类别信息、特征和像素重构信息辅助训练，生成质量更好的图像. 将图像模型分别在 Oxford-102、加利福尼亚理工学院鸟类数据库(CUB)和微软 COCO(MS COCO)数据集上进行验证，使用 Inception Score 评估生成图像的质量和多样性. 结果表明：提出的模型具有一定的效果，在 3 个数据集上的 Inception Score 值分别是 3.54，4.16 和 11.45，相应比堆栈生成对抗网络提高 10.6%，12.4%和 35.5%.

关键词： 文本生成图像；堆栈生成对抗网络；分类；重构；跨模态学习

中图分类号： TP 391.41 **文献标志码：** A **文章编号：** 1000-5013(2019)04-0549-07

Text to Image Model With Classification-Reconstruction Stack Generative Adversarial Networks

CHEN Xinjing, CHEN Duansheng

(College of Computer Science and Technology, Huaqiao University, Xiamen 361021, China)

Abstract: Using the stack generative adversarial networks, we propose classification and reconstruction stack generative adversarial network. We have generated 64 px×64 px resolution images in the stage Ⅰ, then we synthesize 256 px×256 px resolution images in the Stage Ⅱ. In each stage of the text to image, we add the image category information, feature and pixel reconstruction information to assist in generating high-quality images. In this paper, we test and verify the presented model on Oxford-102, Caltech-University of California San Diego Birds (CUB) and Microsoft COCO (MS COCO) datasets, and evaluated the quality and diversity of generated images with Inception Score. The results show that the model proposed in this paper has certain effects, Inception Score on the three datasets were 3.54, 4.16 and 11.45, respectively, which increased by 10.6%, 12.4%, and 35.5% over the stack generative adversarial networks.

Keywords: text to image; stack generative adversarial networks; classification; reconstruction; Cross-modal learning

文本生成图像是结合计算机视觉和自然语言处理两个领域的综合性任务，是一个跨学科、跨模态的交叉性任务. 其输入的是一句或一段文本，输出包含该文本语义信息的图像，具有很大的实用价值. 例如，为不同品种的花卉、鸟类配插图等. 在文本生成图像任务中，不仅需要计算机理解文本语义信息，还要将其转化为像素，是一项极具挑战性的工作，但是随着深度学习日渐火热，在跨模态应用取得了很大

收稿日期： 2018-07-22

通信作者： 陈锻生(1959-)，男，教授，博士，主要从事数字图像处理与模式识别的研究. E-mail：dschen@hqu.edu.cn.

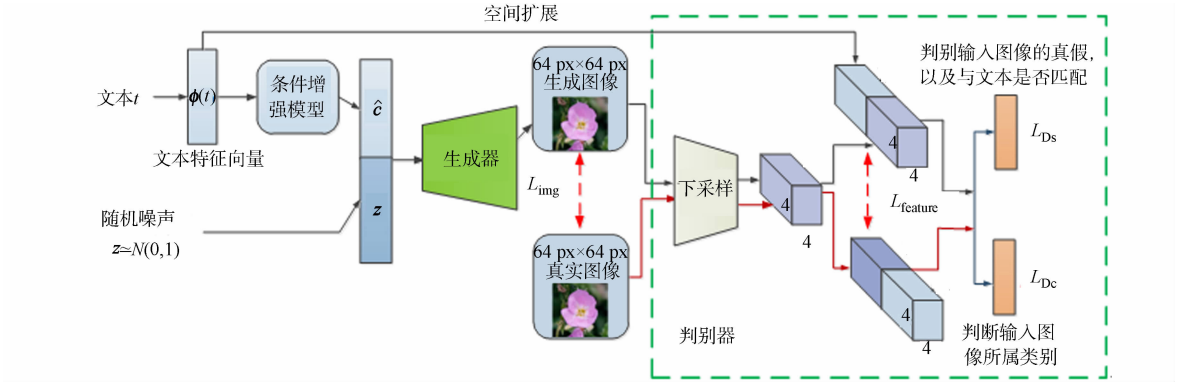
基金项目： 国家自然科学基金资助项目(61502182)；福建省科技计划重点项目(2015H0025)

突破[1-3]. 另外,生成对抗网络(generative adversarial networks,GAN)[4]也为这项任务提供了可能. 学者对于文本生成图像的研究取得了一定的突破[5-13],但是生成图像的颜色和细节处理还有待提高. 为了进一步改善生成图像的质量,本文提出分类重构堆栈生成对抗网络文本生成图像模型(CRStack-GAN).

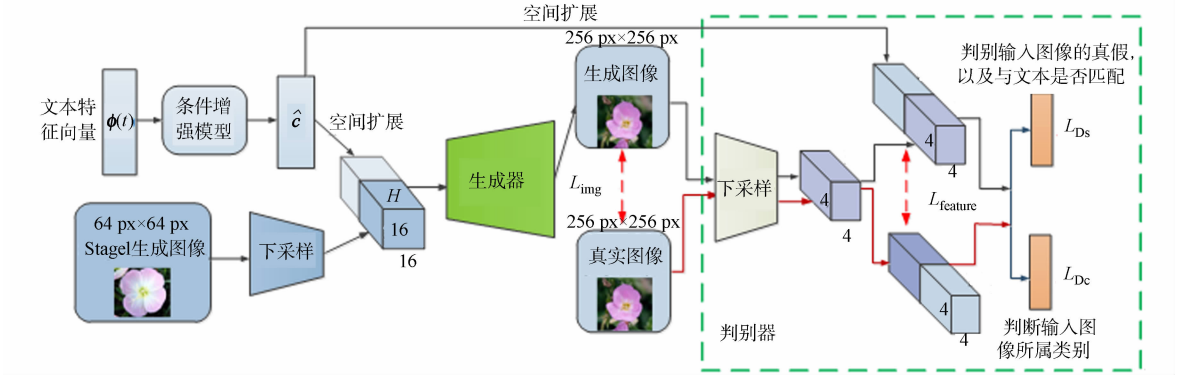
1 分类重构堆栈生成对抗网络

1.1 对抗网络模型结构

在图像生成中,很难直接生成高质量的图像,提出的分类重构堆栈生成对抗网络分两个阶段进行训练,每个阶段的模型框架,如图1所示. 图1中:$\boldsymbol{\phi}(t)$为文本描述t的特征向量;L_{Ds}为判别损失;L_{Dc}为分类损失;$\hat{\boldsymbol{c}}$为条件变量;$\boldsymbol{z}$为随机噪声;H为交叉熵损失;L_{feature}特征重构损失;L_{img}为图像重构损失.



(a) 第Ⅰ阶段



(b) 第Ⅱ阶段

图1 分类重构堆栈生成对抗网络模型结构图

Fig. 1 Architecture of CRStack-GAN

由图1(a)可知:将文本通过条件增强模型后,对生成的向量和噪声向量进行拼接,通过生成器生成64 px×64 px的图像,在判别器中,将64 px×64 px的图像(生成或真实图像)经过下采样后,对得到的张量和文本特征向量进行拼接,经过两个平行的卷积层得到两个不同的概率分布,第一个概率分布判别输入图像的真假,以及与文本是否匹配,第2个概率分布判别输入图像所属类别.

由图1(b)可知:将文本特征向量经过条件增强模型后的张量与第一阶段生成的64 px×64 px的图像经过下采样后得到的张量进行拼接,通过生成器生成256 px×256 px的图像,第Ⅱ阶段的判别器与第Ⅰ阶段相同,只是将输入图像改为256 px×256 px.

图1与文献[9]中的模型图最大的不同点在于判别器的设置,图1在判别器的末尾增加了一个分类器,用于判断输入图像的所属类别,并且在生成器的损失计算中引入像素重构损失L_{img}和L_{feature},提高生成图像的质量.

1.2 条件增强模型

首先,将文本描述表示为特征向量,采用文献[14]预训练的字符级卷积循环神经网络(char-CNN-

RNN)文本编码器，将文本描述表示为特征向量 $\boldsymbol{\phi}(t)$，$\boldsymbol{\phi}(t)\in\mathbf{R}^{N_t}$，由于 $\boldsymbol{\phi}(t)$是一个高维向量($N_t>100$)，在数据量有限的情况下，导致潜在数据流不连续，不利于生成模型的学习. 为了克服这个问题，采用文献[9]提出的条件增强模型，将向量 $\boldsymbol{\phi}(t)$转化为低维的条件变量 $\hat{\boldsymbol{c}}$，$\hat{\boldsymbol{c}}\in\mathbf{R}^{N_c}$.

在条件增强模型中，将向量 $\boldsymbol{\phi}(t)$通过全连接层得到均值 $\boldsymbol{\mu}(t)$和协方差矩阵 $\boldsymbol{\Sigma}(t)$，然后，从单位高斯分布 $N(0,\boldsymbol{I})$中随机采样 $\boldsymbol{\varepsilon}$，则条件变量 $\hat{\boldsymbol{c}}$ 为

$$\hat{\boldsymbol{c}}=\boldsymbol{\mu}(t)+\boldsymbol{\Sigma}(t)\odot\boldsymbol{\varepsilon}. \tag{1}$$

式(1)中：$\odot$表示矩阵元素对应相乘. 为了增强平滑度避免过拟合，生成器在训练的过程中增加以下损失函数，即

$$L_{\mathrm{KL}}\leftarrow D_{\mathrm{KL}}(N(\boldsymbol{\mu}(t);\boldsymbol{\Sigma})\parallel N(0;\boldsymbol{I})). \tag{2}$$

式(2)中：L_{KL}为 KL 的重构损失；D_{KL}表示 KL 散度.

1.3　第Ⅰ阶段

在第Ⅰ阶段，将文本特征向量经过条件增强模型后得到的条件变量 $\hat{\boldsymbol{c}}$ 和随机噪声 $\boldsymbol{z}$ 进行拼接，$\boldsymbol{z}\in\mathbf{R}^{N_z}$，通过生成器生成 64 px×64 px 图像 $\boldsymbol{I}_{\mathrm{f0}}$. 在具体生成器中，将 $\hat{\boldsymbol{c}}$ 和 $\boldsymbol{z}$ 拼接后，通过全连接层将其变成大小为 $4\times4\times N_{\mathrm{g}}$ 的张量，然后，将其通过上采样(反卷积)操作生成 64 px×64 px 的图像，试验中，$N_t=1\ 024$，$N_c=128$，$N_z=128$.

在第Ⅰ阶段的判别网络中，输入 64 px×64 px 的图像和文本，得到图像源和图像文本匹配度的概率分布 D_{s0}和图像类别标签的概率分布 D_{c0}. 首先，将文本特征向量 $\boldsymbol{\phi}(t)$通过全连接层将其转化为潜在变量 $\boldsymbol{l}_{\mathrm{r0}}$，$\boldsymbol{l}_{\mathrm{r0}}\in\mathbf{R}^{N_l}$，将 $\boldsymbol{l}_{\mathrm{r0}}$通过空间扩展后变成大小为 $4\times4\times N_l$ 的张量 $\boldsymbol{N}$. 然后，将 64 px×64 px 的图像 $\boldsymbol{I}_0$ 通过下采样后变成大小为 $4\times4\times F$ 的张量 $\boldsymbol{M}$. 将 $\boldsymbol{M}$ 和 $\boldsymbol{N}$ 在 F 通道上进行拼接后，得到中间层特征张量 $\boldsymbol{Q}$，其大小为 $4\times4\times(F+N_l)$. 最后，将 $\boldsymbol{Q}$ 分别传入两个卷积核个数为 1 和 c 的卷积层中. 卷积核为 1 的卷积层产生图像源和图像文本匹配度的概率分布 D_{s0}，卷积核为 c 的卷积层产生图像类别标签的概率分布 D_{c0}.

1.4　第Ⅱ阶段

在第Ⅱ阶段，输入文本和第Ⅰ阶段生成的 64 px×64 px 的图像 $\boldsymbol{I}_{\mathrm{f0}}$，输出 256 px×256 px 的图像 $\boldsymbol{I}_{\mathrm{f1}}$. 首先，将第Ⅰ阶段生成的 64 px×64 px 的图像 $\boldsymbol{I}_{\mathrm{f0}}$通过卷积下采样后变成大小为 $16\times16\times H$ 的 $\boldsymbol{T}$. 将文本特征向量 $\boldsymbol{\phi}(t)$通过条件增强模型转化为 N_{c} 维的 $\hat{\boldsymbol{c}}$，将 $\hat{\boldsymbol{c}}$ 通过空间扩展变成大小为 $16\times16\times N_c$ 的 $\boldsymbol{P}$. 将 $\boldsymbol{T}$ 和 $\boldsymbol{P}$ 在 H 通道上进行拼接后，变成大小为 $16\times16\times(H+N_c)$的张量 $\boldsymbol{B}$. 最后，将拼接后张量 $\boldsymbol{B}$ 通过上采样生成 256 px×256 px 的图像 $\boldsymbol{I}_{\mathrm{f1}}$.

第Ⅱ阶段的判别器和第Ⅰ阶段类似，将文本特征向量 $\boldsymbol{\phi}(t)$通过全连接层转化为潜在变量 $\boldsymbol{l}_{\mathrm{r1}}$，$\boldsymbol{l}_{\mathrm{r1}}\in\mathbf{R}^{N_l}$，将 $\boldsymbol{l}_{\mathrm{r1}}$通过空间扩展变成大小为 $4\times4\times N_l$ 的张量 $\boldsymbol{N}_1$，将 256 px×256 px 的图像 $\boldsymbol{I}_1$ 通过下采样变成大小为 $4\times4\times F$ 的张量 $\boldsymbol{M}_1$，将 $\boldsymbol{M}_1$ 和 $\boldsymbol{N}_1$ 在 F 通道上进行拼接，得到中间层特征张量 $\boldsymbol{Q}_1$. 最后，将 $\boldsymbol{Q}_1$ 分别通过两个卷积层，得到图像源和图像文本匹配度的概率分布 D_{s1}，以及图像类别标签的概率分布 D_{c1}.

1.5　训练机制

第Ⅰ，Ⅱ阶段采用同样的训练方式，在每个阶段的训练中，分别将图像文本对{($\boldsymbol{I}_{\mathrm{r}}$，$\boldsymbol{l}_{\mathrm{r}}$)，($\boldsymbol{I}_{\mathrm{f}}$，$\boldsymbol{l}_{\mathrm{r}}$)，($\boldsymbol{I}_{\mathrm{w}}$，$\boldsymbol{l}_{\mathrm{r}}$)}传入判别器中，其中，$\boldsymbol{I}_{\mathrm{r}}$ 为与文本 $\boldsymbol{l}_{\mathrm{r}}$ 相匹配的真实图像；$\boldsymbol{I}_{\mathrm{f}}$ 为生成器生成的虚假图像；$\boldsymbol{I}_{\mathrm{w}}$ 为与文本 $\boldsymbol{l}_{\mathrm{r}}$ 不匹配的错误图像，则判别损失 L_{Ds}和分类损失 L_{Dc}分别为

$$L_{\mathrm{Ds}}\leftarrow H(\mathrm{Ds}(\boldsymbol{I}_{\mathrm{r}},\boldsymbol{l}_{\mathrm{r}}),1)+(H(\mathrm{Ds}(\boldsymbol{I}_{\mathrm{f}},\boldsymbol{l}_{\mathrm{r}}),0)+H(\mathrm{Ds}(\boldsymbol{I}_{\mathrm{w}},\boldsymbol{l}_{\mathrm{r}}),0))\times0.5, \tag{3}$$

$$L_{\mathrm{Dc}}\leftarrow H(\mathrm{Dc}(\boldsymbol{I}_{\mathrm{r}},\boldsymbol{l}_{\mathrm{r}}),C_{\mathrm{r}})+(H(\mathrm{Dc}(\boldsymbol{I}_{\mathrm{f}},\boldsymbol{l}_{\mathrm{r}}),C_{\mathrm{r}})+H(\mathrm{Dc}(\boldsymbol{I}_{\mathrm{w}},\boldsymbol{l}_{\mathrm{r}}),C_{\mathrm{w}}))\times0.5. \tag{4}$$

式(3)，(4)中：H 表示交叉熵损失；C_{r} 和 C_{w} 分别为图像 $\boldsymbol{l}_{\mathrm{r}}$ 和 $\boldsymbol{I}_{\mathrm{w}}$ 的类别标签.

判别器的损失函数为

$$L_{\mathrm{D}}=L_{\mathrm{Ds}}+L_{\mathrm{Dc}}. \tag{5}$$

在生成器的训练中，$L_{\mathrm{Gs}}\leftarrow H(\mathrm{Ds}(\boldsymbol{I}_{\mathrm{f}},\boldsymbol{l}_{\mathrm{r}}),1)$，$L_{\mathrm{Gc}}\leftarrow H(\mathrm{Dc}(\boldsymbol{I}_{\mathrm{f}},\boldsymbol{l}_{\mathrm{r}}),C_{\mathrm{r}})$，则生成器的损失函数为

$$L_{\mathrm{G0}}=L_{\mathrm{Gs}}+L_{\mathrm{Gc}}+\lambda_0 L_{\mathrm{kl}}. \tag{6}$$

式(6)中：L_{G0}为单纯地加入图像的分类损失；λ_0 为超参数.

为了使生成器更好地收敛，以及使生成模型不太偏离真实样本，在模型生成器的损失函数中，加入

特征重构误差和像素重构误差. 其中:特征重构误差 $L_{\text{feature}}=\| f_{\text{D}}(\boldsymbol{I}_{\text{f}},\boldsymbol{l}_{\text{r}})-f_{\text{D}}(\boldsymbol{I}_{\text{r}},\boldsymbol{l}_{\text{r}})\|_2^2$. 其中,$f_{\text{D}}(\boldsymbol{I}_{\text{f}},\boldsymbol{l}_{\text{r}})$和 $f_{\text{D}}(\boldsymbol{I}_{\text{r}}),\boldsymbol{l}_{\text{r}}$ 分别表示图像文本对$\{(\boldsymbol{I}_{\text{f}},\boldsymbol{l}_{\text{r}}),(\boldsymbol{I}_{\text{r}},\boldsymbol{l}_{\text{r}})\}$在判别器中得到的中间层特征 $\boldsymbol{Q}$. 像素重构误差为 $L_{\text{img}}=\| \boldsymbol{I}_{\text{f}}-\boldsymbol{l}_{\text{r}}\|_2^2$,则生成器的损失函数变为

$$L_{\text{G1}}=L_{\text{Gs}}+L_{\text{Gc}}+\lambda_0 L_{\text{kl}}+\lambda_1 L_{\text{feature}}+\lambda_2 L_{\text{img}}. \tag{7}$$

式(7)中:$\lambda_0=2,\lambda_1=1,\lambda_2=1$;$L_{\text{G1}}$表示加入图像的类别信息以及特征和像素的重构损失.

2 实验结果与分析

2.1 数据集及评价指标

为了验证模型的有效性,分别在 Oxford-102[6],CUB[7]和 MS COCO[8]数据集上进行实验,数据集如表1所示.

表1 实验数据

Tab. 1 Experimental data

数据集	Oxford-102[6]		CUB[7]		MS COCO[8]	
	训练集	测试集	训练集	测试集	训练集	测试集
图像数量	7 034	1 155	8 855	2 933	80 000	40 000
图像描述数量	10	10	10	10	5	5

实验采用文献[9]中同样的评价指标 Inception Score[15],其计算公式为

$$I=\exp(E_x D_{\text{KL}}(p(y\mid x)\| p(y))). \tag{8}$$

式(8)中:x 为生成样本;y 为 Inception model 预测的标签.

一个好的生成模型应该生成多样且有意义的图像,因此,边缘分布 $p(y)$和条件分布 $p(y|x)$的 KL 散度应该越大越好.

2.2 模型的选择

为了验证节1.5提出的特征和像素重构损失的有效性,在 Oxford-102 数据集上进行对比实验,生成效果图,如图2所示.

图2 生成器的损失函数在 Oxford-102 测试集上生成图像示例

Fig. 2 Example results conditioned on text descriptions from Oxford-102 test set

由图2可知:在 Stack-GAN 的基础上,加入分类损失 L_{Gc}后生成的图像颜色处理比 Stack-GAN 好,但是随着训练的不断进行,容易远离真实图像. 在加入了特征和像素重构损失后,生成的图像更加清晰,

细节部分也更细腻真实.

生成器不同损失在 Oxford-102[6] 测试集上的 Inception Score 值，如表 2 所示. 表 2 中：$L_G = L_{Gs} + \lambda_0 L_{kl}$.

表 2　生成器不同损失在 Oxford-102[6] 测试集上的 Inception Score 值

Tab. 2　Inception scores by different generator loss on Oxford-102[6] test sets

模型	损失函数	I(Oxford-102)
Stack-GAN	$L_G = L_{Gs} + \lambda_0 L_{kl}$	3.20±0.01
CRStack-GAN	$L_{G0} = L_{Gs} + L_{Gc} + \lambda_0 L_{kl}$	3.44±0.04
	$L_{G1} = L_{Gs} + L_{Gc} + \lambda_0 L_{kl} + \lambda_1 L_{feature} + \lambda_2 L_{img}$	3.54±0.03

2.3　模型对比

生成图像示例，如图 2～4 所示. 由图 2～4 可知：提出的 CRStack-GAN 比 StackGAN 生成图像的颜色及细节处理更加细腻，图像也更加清晰.

图 3　在 CUB 测试集上生成图像示例

Fig. 3　Example results by conditionedon text descriptions from CUB test set

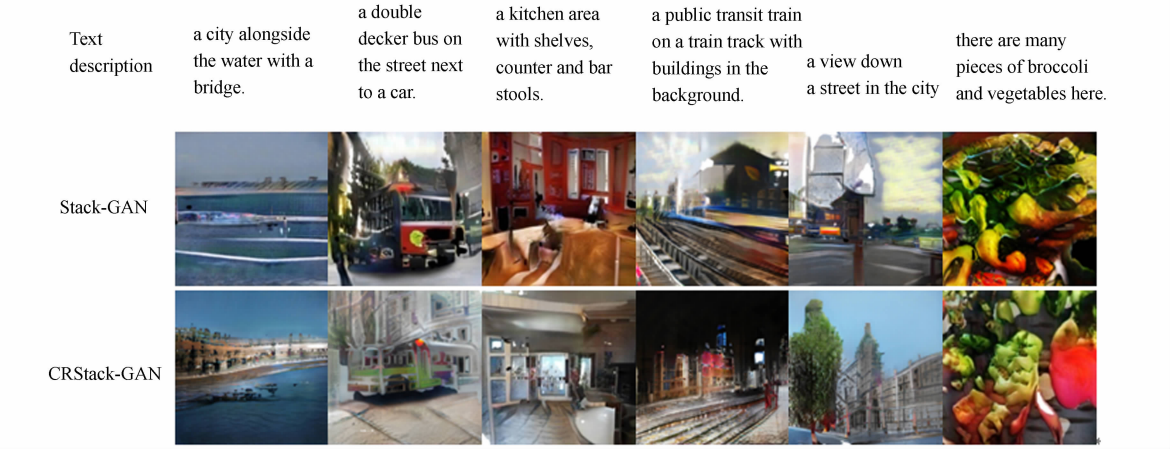


图 4　Stack-GAN[9] 和提出的 CRStack-GAN 在 MS COCO 测试集上生成图像示例

Fig. 4　Example results by Stack-GAN[9] and our CRStack-GAN conditioned on text descriptions from MS COCO test set

为了验证模型的有效性，实验在 Oxford-102[6]，CUB[7] 和 MS COCO[8] 数据集上进行实验，同文献[9]一样，分别在 Oxford-102，CUB 测试集上随机生成 3 000 张图像，使用文献[9]在这两个数据集上预训练的 Inception model 进行评估，在 MS COCO 测试集上生成 4 000 张图像，使用文献[15]在 ImageNet 数据集上预训练的 Inception model 进行评估. 实验结果，如表 3 所示. 由表 3 可知：提出的模型在 3

个数据集上的 Inception Score 值分别为 3.54,4.16 和 11.45,高于最近提出的很多模型[10-13],相应地比 Stack-GAN 提高了 10.6%,12.4%和 35.5%.

表 3 各模型分别在 Oxford-102,CUB 和 MS COCO 测试集上的 Inception Score 值

Tab. 3 Inception scores by models on Oxford-102, CUB and MS COCO test sets

数据集	I(Stack-GAN)	I(Stack-GAN-V2)	I(TAC-GAN)	I(PPGN)	I(ChatPainter)	I(CRStack-GAN)
Oxford-102	3.20±0.01	—	3.45±0.05	—	—	3.54±0.03
CUB	3.70±0.04	4.04±0.05	—	—	—	4.16±0.03
MSCOCO	8.45±0.03	—	—	9.58±0.21	9.74±0.02	11.45±0.18

2.4 风格迁移

为了进一步测试输出图像对输入句子变化的敏感程度,在 Oxford-102 测试集上通过改变文字描述中最受关注的单词生成相应图像,如图 5 所示. 分别改变句子图 5(a)中的颜色词汇,生成相应图像,如图 5(b),(c)和(d)所示. 由图 5 可知:生成的图像根据输入句子的变化而变化,模型可以捕捉文本描述中的细微语义差异.

(a) 花瓣是粉红色的,中心是黄色的

(b) 花瓣是红色的,中心是黄色的

(c)花瓣是蓝色的,中心是白色的

(d) 花瓣是白色的,中心是黄色的

图 5 CRStack-GAN 在 Oxford-102 测试集上通过改变文本描述生成的图像示例

Fig. 5 Example results of CRStack-GAN model trained on Oxford-102 while changing some most attended words in text descriptions

3 结论

针对文本生成图像任务,提出一种分类重构堆栈生成对抗网络文本生成图像模型,沿用堆栈生成对抗网络的思想,在其基础上增加了图像类别信息、特征重构和像素重构信息辅助训练,提高了生成图像的质量.

通过实验对比,提出的模型比堆栈生成对抗网络分别在 Oxford-102,CUB 和 MS COCO 数据集上的 Inception Score 提高了 10.6%,12.4%和 35.5%. 生成的图像颜色和细节部分处理的更加细腻,也能够很好地捕捉到文本描述中的细微语义差异. 针对场景比较复杂的数据集,结合视觉对话语料[16]进一步提高生成图像的效果是以后的研究工作.

参考文献：

[1] XU K，BA J，KIROS R，*et al*. Show，attend and tell：Neural image caption generation with visual attention[C]// International Conference on Machine Learning. Lille：[s. n.]，2015：2048-2057.

[2] 邹辉杜，吉祥翟，传敏，等. 深度学习与一致性表示空间学习的跨媒体检索[J]. 华侨大学学报(自然科学版)，2018，39(1)：127-132. DOI：10. 11830/ISSN. 1000-5013. 201508047.

[3] WEI Yunchao，ZHAO Yao，LU Canyi，*et al*. Cross-modal retrieval with CNN visual features：A new baseline[J]. IEEE Transactions on Cybernetics，2017，47(2)：449-460. DOI：10. 1109/TCYB. 2016. 2519449.

[4] GOODFELLOW I，POUGET-ABADIE J，MIRZA M，*et al*. Generative adversarial nets[C]// Advances in Neural Information Processing Dystems. Montreal：[s. n.]，2014：2672-2680.

[5] REED S. Generative adversarial text to image synthesis[J]. International Machine Learning Society，2016(48)：1060-1069.

[6] NILSBACK M E，ZISSERMAN A. Automated flower classification over a large number of classes[C]// Conference on Computer Vision，Graphics and Image Processing. Washington：IEEE Press，2008：722-729. DOI：10. 1109/ICVGIP. 2008. 47.

[7] WAH C，BRANSON S，WELINDER P，*et al*. Caltech-UCSD birds 200[EB/OL]. [2011-10-26][2018-06-15]. http：// www. vision. caltech. edu/visipedia/CUB-200. html.

[8] LIN Tsungyi，MAIRE M，BELONGIE S，*et al*. Microsoft COCO：Common objects in context[C]// European Conference on Computer Vision. Zurich：[s. n.]，2014：740-755.

[9] ZHANG Han，XU Tao，LI Hongsheng，*et al*. Stackgan：Text to photo-realistic image synthesis with stacked generative adversarial networks[J]. IEEE International Conference on Computer Vision，2017，2(3)：5908-5916. DOI：10. 1109/ICCV. 2017. 629.

[10] ZHANG Han，XU Tao，LI Hongsheng，*et al*. Stackgan＋＋：Realistic image synthesis with stacked generative adversarial networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2017(99)：1. DOI：10. 1109/TPAMI. 2018. 2856256.

[11] AYUSHMAN D. TAC-GAN-text conditioned auxiliary classifier generative adversarial network[J/OL]. [2017-03-26][2018-07-10]. https：// arxiv. org/abs/1703. 06412.

[12] NGUYEN A. Plug and play generative networks：Conditional iterative generation of images in latent space[J]. IEEE Conference on Computer Vision and Pattern Recognition，2017(21)：3510-3520. DOI：10. 1109/CVPR. 2017. 374.

[13] SHIKHAR S. ChatPainter：Improving text to image generation using dialogue[J/OL]. [2018-02-22][2018-06-12]. https：// arxiv. org/abs/1802. 08216.

[14] REED S，AKATA Z，LEE H，*et al*. Learning deep representations of fine-grained visual descriptions[C]// Proceedings of the IEEE Computer Vision and Pattern Recognition. Las Vegas：IEEE Press，2016：49-58. DOI：10. 1109/CVPR. 2016. 13.

[15] SALIMANS T，GOODFELLOW I，ZAREMBA W，*et al*. Improved techniques for training gans[C]// Advances in Neural Information Processing Systems. Barcelona：[s. n.]，2016：2234-2242.

[16] DAS A，KOTTUR S，GUPTA K，*et al*. Visual dialog[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas：IEEE Press，2017：1080-1089. DOI：10. 1109/TPAMI. 2018. 2828437.

（责任编辑：陈志贤　　英文审校：吴逢铁）