

多维度注意力和语义再生的文本生成图像模型

庄兴旺,丁岳伟

(上海理工大学 光电信息与计算机工程学院,上海 200093)

摘要: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,从给定的文本描述生成图像有两个目标:视觉真实性和语义一致性。虽然在使用生成对抗网络(GAN)生成高质量和视觉逼真的图像方面取得了显著进展,但确保文本描述和视觉内容之间的语义一致性仍然是非常具有挑战性的。目前的方法由于文本和图像形式的多样性,仅在单词级别使用注意力并不能确保全局语义的一致性。因此,在MirrorGAN的基础上提出了一种改进的多维度的注意力协同模块(MCAM)和语义文本再生模块(STRM)来解决这些问题。MCAM使用了更为先进的BERT模型来进行文本处理,STRM用于从生成的图像中重新生成文本描述,该图像在语义上与给定的文本描述对齐,使生成的图像更加贴合语义。最后,形成了基于多维度注意力以及语义文本再生的生成对抗网络模型(MirrorGAN++)。通过对两个公共基准数据集的深入实验,证明了MirrorGAN++优于其他方法。

关键词:文本生成图像;生成对抗网络;语义一致;注意力;语义文本再生

中图分类号:TP391.41

文献标识码:A

文章编号:1673-629X(2020)12-0027-07

doi:10.3969/j.issn.1673-629X.2020.12.005

Text-to-image Model by Multidimensional Attention and Semantic Regeneration

ZHUANG Xing-wang, DING Yue-wei

(School of Optoelectrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China)

Abstract: Text-to-image is a comprehensive task combining computer vision and natural language processing. Generating an image from a given text description has two goals: visual realism and semantic consistency. Although significant progress has been made in generating high-quality and visually realistic images using generative adversarial networks, guaranteeing semantic consistency between the text description and visual content remains challenging. The current approaches only using word-level attention cannot ensure global semantic consistency due to the diverse nature of both the text and image modalities. Therefore, we propose an improved multidimensional collaborative attentive module (MCAM) and semantic text regeneration module (STRM) based on MirrorGAN to solve these problems. MCAM uses a more advanced BERT model for text processing, and STRM is used to regenerate the text description from the generated image. The image is semantically aligned with the given text description, making the generated image more suitable for the semantics. Finally, a generative adversarial network model based on multi-dimensional attention and semantic text regeneration (MirrorGAN++) is formed. Thorough experiments on two public benchmark datasets demonstrate the superiority of MirrorGAN++ over other representative state-of-the-art methods.

Key words: text-to-image; GAN; semantic consistency; attention; semantic text regeneration

0 引言

文本生成图像(text to image, T2I)是指生成与给定文本描述匹配的视觉真实图像。由于其在许多应用领域的巨大潜力, T2I已经成为自然语言处理和计算机视觉的一个重要研究领域。虽然在使用生成对抗

网络(GAN)生成视觉逼真的图像方面取得了重大进展,如文献[1-8]所示,但确保生成的图像与输入文本的语义对齐仍然具有很大的挑战性。

与基本的图像生成问题相比, T2I是以文本描述为条件的,而不是仅从噪声开始。利用GAN^[9]的强大

收稿日期:2019-12-24

修回日期:2020-04-24

基金项目:上海重点科技攻关项目(16DZ1203603);上海市工程中心建设项目(GCZX14014)

作者简介:庄兴旺(1996-),男,硕士,研究方向为自然语言识别、图像识别、深度学习;丁岳伟,教授,硕导,研究方向为计算机网络及应用、信息安全、电子政务、软件工程和CMM/CMMI等。

功能,提出了不同的 T2I 方法来生成视觉逼真的文本图像。例如,Reed 等人提出了一种解决文本到图像合成问题的方法,即找到文本描述的视觉识别表示,并利用该表示生成真实的图像^[10]。Zhang 等人提出了在两个独立的阶段生成图像的 Stackgan^[2]。Hong 等人提出了从输入文本中提取语义布局,然后将其转换为图像生成器,以指导生成过程^[5]。Zhang 等人提出在网络层次结构中引入了伴随的层次嵌套对抗性目标,它规范了中间层的表示,并帮助生成器训练来捕获复杂的图像统计信息^[3]。这些方法都仅利用鉴别器来区分。然而,由于文本和图像之间域的差异,当单独依赖于这样一个鉴别器时,很难对语义一致性进行建模。最近,人们利用注意力机制^[4]来解决这一问题,它引导生成器在生成不同的图像区域时关注于不同的单词。然而,由于文本和图像形式的多样性,仅在单词级别使用注意力并不能确保全局语义的一致性。T2I 生成可以看作是图像字幕(或图像到文本生成, I2T)的逆问题^[11-12]。如果 T2I 生成的图像在语义上与给定的文本描述一致,那么 T2I 重新描述的语义应该与给定的文本描述完全相同。基于这一观察,该文在 MirrorGAN 的基础上提出了一种新的文本到图像到文本的模型 MirrorGAN++ 来改进 T2I 的生成。MirrorGAN++ 有两个模块:MCAM 和 STRM。MCAM 是多维度的注意力协同模块,利用单词级别的注意和全局句子级别的注意逐步增强生成图像的多样性和语义一致性。STRM 是语义文本再生模块,它可从最后生成的图像重新生成文本描述,在语义上与给定的文本描述对齐。对两个公共基准数据集进行的深入实验表明,在视觉真实性和语义一致性方面, MirrorGAN++ 优于其他方法。

1 相关工作

CycleGAN^[13-15]可以让两个领域的图片互相转化^[16-17],传统的 GAN 是单向生成,而 CycleGAN 是互相生成。MirrorGAN++ 部分灵感来自 CycleGAN,但有两个主要区别:

(1) MirrorGAN++ 专门解决 T2I 问题,而不是图像到图像的转化。文本和图像之间的跨媒体域间隙可能比具有不同属性(例如样式)的图像之间的间隙大得多。此外,每个域中存在的不同语义使得保持跨域语义一致性变得更加困难。

(2) MirrorGAN++ 通过使用成对的文本图像数据进行监督学习,而不是从不成对的图像数据进行训练。此外,为了体现通过重新描述学习 T2I 生成的思想,使用基于 CE 的重构损失来规范重新描述的文本的语义一致性,这与 CycleGAN 中的 L1 循环一致性损失不同,后者处理视觉相似性。该模型是基于 MirrorGAN 的,然而,由于 MirrorGAN 对输入的文本是用 RNN 来处理,使用 RNN 生成的词嵌入和句嵌入表现没有 BERT 生成的好,所以选择用 BERT 来生成词嵌入和句嵌入,这样可以更好地提升模型生成图片的质量。对于语义文本再生模块,选用更加先进的图像字幕模型来更好地提升模型的语义一致性。

2 模型实现

如图 1 所示, MirrorGAN++ 集成了 T2I 和 I2T 模块,它利用了重新描述学习 T2I 生成的思想。生成图像后, MirrorGAN++ 会重新生成其描述,从而将其语义与给定的文本描述对齐。从技术上讲, MirrorGAN++ 由两个模块组成:MCAM 和 STRM。模型的细节将在下面介绍。

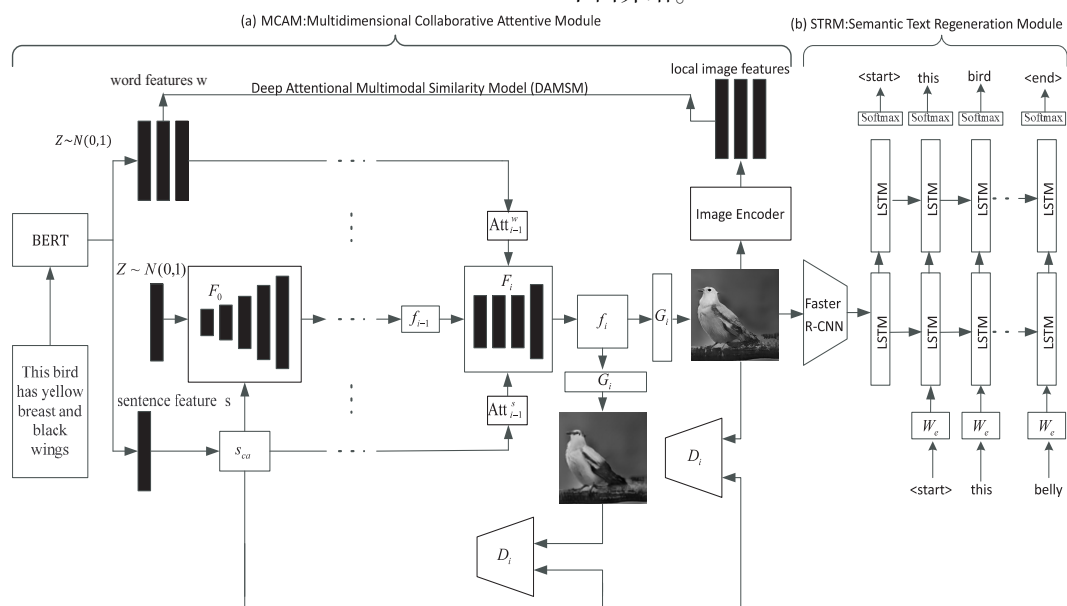


图 1 用于文本到图像生成的模型 MirrorGAN++ 示意图

其中, \mathbf{W}_e 表示一个词嵌入矩阵, 它将文字特征映射到视觉特征空间, \prod_i 是输入单词的独热编码。

对于 Attention LSTM 的输出 h_t^1 , 在每个时间步 t 时, 为每一个图像特征 v_i 生成标准化注意权重 $\alpha_{i,t}$, 如下所示:

$$\begin{aligned}\alpha_{i,t} &= w_a^T \tanh(W_{va} v_i + W_{ha} h_t^1) \\ \alpha_i &= \text{softmax}(a_i)\end{aligned}\quad (8)$$

其中, $W_{va} \in R^{H \times M}$, $W_{ha} \in R^{H \times M}$, $w_a \in R^H$ 是学习参数。

\hat{v}_i 作为 Language LSTM 输入的图像特征是所有输入特征的凸组合:

$$\hat{v}_i = \sum_{i=1}^K \alpha_{i,t} v_i \quad (9)$$

Language LSTM 的输入包括图像特征和 Attention LSTM 的输出, 由下式给出:

$$x_t^2 = [\hat{v}_i, h_t^1] \quad (10)$$

词的概率分布为:

$$p_t = \prod_{i=1}^T \text{softmax}(W_p h_t^2 + b_p) \quad (11)$$

其中, $W_p \in R^{|\Sigma| \times M}$, $b_p \in R^{|\Sigma|}$ 是学习的权重和偏差。

2.3 目标函数

根据普遍的做法, 首先采用了两种对抗性损失: 视觉和真实的对抗性损失和文本图像对语义一致性的对抗性损失, 定义如下:

在模型训练的每个阶段, 生成器 G 和鉴别器 D 交替训练。特别是生成器 G_i 在第 i^{th} 阶段通过最小化损失进行训练, 如下所示:

$$\begin{aligned}L_{G_i} &= -\frac{1}{2} E_{I_i \sim p_i} [\log(D_i(I_i))] - \\ &\quad \frac{1}{2} E_{I_i \sim p_i} [\log(D_i(I_i, s))] \end{aligned}\quad (12)$$

其中, I_i 是在第 i 阶段的分布 p_{I_i} 中采样生成的图像。第一项是视觉和真实的对抗性损失, 用来区分图像是真实的还是虚假的。第二项是文本图像对语义一致性的对抗性损失, 用来确定图像和句子语义是否是一致的。

进一步提出了一个基于 CE 的文本语义重构损失, 以在 STRM 的重新描述和给定的文本描述之间对齐语义。从数学上讲, 这种损失可以表示为:

$$L_{\text{strm}} = - \sum_{i=0}^{L-1} \log p_i(T_i) \quad (13)$$

值得注意的是, 在 STRM 预训练期间, 也使用了 L_{strm} 。当训练 G_i 时, 从 L_{strm} 到 G_i 的梯度通过 STRM 反向传播, 其网络权重保持不变。

使用 AttnGAN 的 DAMSM 损失^[4] L_{DAMSM} 来测量图像和文本描述之间的匹配度。 L_{DAMSM} 使生成的图像更

好地依赖于文本描述。

生成器的最终目标函数定义为:

$$L_G = \sum_{i=0}^{m-1} L_{G_i} + \lambda_1 L_{\text{strm}} + \lambda_2 L_{\text{DAMSM}} \quad (14)$$

其中, λ_1 和 λ_2 是调节文本语义重构损失和 DAMSM 损失的相应权重。

鉴别器 D_i 被交替训练, 以避免被生成器骗过去, 将输入分为实输入和假输入。与生成器类似, 鉴别器的目标函数包括视觉和真实的对抗性损失和文本图像对语义一致性的对抗性损失。在数学上, 它可以定义为:

$$\begin{aligned}L_{D_i} &= -\frac{1}{2} E_{I_i^{\text{GT}} \sim p_{I_i^{\text{GT}}}} [\log(D_i(I_i^{\text{GT}}))] - \\ &\quad \frac{1}{2} E_{I_i \sim p_{I_i}} [\log(1 - D_i(I_i))] - \\ &\quad \frac{1}{2} E_{I_i^{\text{GT}} \sim p_{I_i^{\text{GT}}}} [\log(D_i(I_i^{\text{GT}}, s))] - \\ &\quad \frac{1}{2} E_{I_i \sim p_{I_i}} [\log(1 - D_i(I_i, s))] \end{aligned}\quad (15)$$

其中, I_i^{GT} 来自第 i^{th} 阶段的真实的图像分布 $p_{I_i^{\text{GT}}}$ 。

鉴别器的最终目标函数定义为:

$$L_D = \sum_{i=0}^{m-1} L_{D_i} \quad (16)$$

3 实验研究

在这一部分中, 进行了大量的实验, 以评估提出的模型。首先将 MirrorGAN++ 与之前的 T2I 方法, 如 GAN-INT-CLS^[10]、StackGAN^[2]、StackGAN++^[22]、PPGN^[23]、AttnGAN^[4] 和 MirrorGAN^[8] 进行比较。然后, 介绍了对 MirrorGAN++ 中 MCAM 和 STRM 的关键部分的消融研究。

3.1 实验方案

3.1.1 数据集

在两个常用数据集上 (CUB bird 数据集^[24] 和 MS COCO 数据集^[25]) 对模型进行评估。CUB bird 数据集包含 8 855 个训练图像和 2 933 个测试图像, 包含 200 个类别, 每个鸟图像有 10 个文本描述。COCO 数据集包含 82 783 个训练图像和 40 504 个验证图像, 每个图像有 5 个文本描述。两个数据集使用文献[2, 4]中的方法进行预处理。

3.1.2 评价指标

首先, Inception Score^[26] 被用来衡量生成图像的客观性和多样性。使用文献[2]提供的两个模型来计算分数。然后, 使用文献[4]中引入的 R-precision 来评估生成的图像与其相应文本描述之间的视觉语义一致性。对于每个生成的图像, 使用其真实的文本描述和从测试集中随机选择的 99 个不匹配描述来形成文本

描述池。然后,在计算 R-precision 之前,计算了池中每个描述的图像特征和文本特征之间的余弦相似性。R-precision 的值越高表示生成的图像和输入文本之间的视觉语义一致性越高。Inception Score 和 R-precision 按文献[2,4]计算。

3.1.3 实现细节

MirrorGAN++总共有三个生成器,最后两个生器采用 MCAM,如式(3)所示。逐步生成 64×64 、 128×128 、 256×256 像素的图像。使用预训练的 BERT^[19] 计算文本描述中的语义嵌入。词嵌入 D 的维数是 256。句子长度 L 是 18。视觉嵌入的维度 M_i 设置为 32。视觉特征尺寸为 $N_i = q_i \times q_i$,三个阶段的 q_i 分别为 64、128 和 256。增广后的句嵌入 D' 的维度设置为 100。文本语义重构损失的损失权重 λ_1 设置为 20, DAMSM 损失权重 λ_2 在 CUB bird 数据集中设置为 5,在 MS COCO 数据集中设置为 50。

3.2 实验结果与分析

在本节中,将定量和定性地与其他方法进行比较。首先,对 MirrorGAN++使用 CUB bird 和 COCO 数据集的 Inception Score 和 R-precision 与之前的文本生成图像方法^[2,4,8,10,22-23] 进行比较。然后,对 MirrorGAN++和之前的方法进行了主观视觉比较,以验证 MirrorGAN++的有效性。

3.2.1 定量结果

将 MirrorGAN++与 CUB 和 COCO 测试数据集上的其他模型进行了比较,结果见表 1 和表 2。

表 1 不同模型在 CUB 和 COCO 数据集下的 IS (越高越好)

模型	CUB	COCO
GAN-INT-CLS	2.88±0.04	7.88±0.07
StackGAN	3.70±0.04	8.45±0.03
StackGAN++	3.82±0.06	/
PPGN	/	9.58±0.21
AttnGAN	4.36±0.03	25.89±0.07
MirrorGAN	4.56±0.05	26.47±0.41
MirrorGAN++	4.82±0.07	31.49±0.13

如表 1 所示, MirrorGAN++模型在 CUB 数据集上 IS 达到了 4.82, 远优于其他方法。与 MirrorGAN 相比, MirrorGAN++将 CUB 数据集的 IS 从 4.56 提高到 4.82 (提高 5.70%), 将 COCO 数据集的 IS 从 26.47 提高到 31.49 (提高 18.96%)。实验结果表明, MirrorGAN++模型生成的图像质量更好。

如表 2 所示, MirrorGAN++与 MirrorGAN 相比将 CUB 数据集的 RI 提高了 0.91%, COCO 数据集的 RI 提高了 1.42%。较高的 RI 表明, MirrorGAN++生成的图像和输入文本之间的视觉语义一致性较高, 进一步

证明了所采用的语义文本再生模块的有效性。

表 2 不同模型在 CUB 和 COCO 数据集下的 RI (越高越好)

模型	CUB	COCO
AttnGAN	67.82±4.43	85.47±3.69
MirrorGAN	70.52±2.28	87.59±1.42
MirrorGAN++	71.43±1.31	89.01±0.94

3.2.2 定性结果

对于定性评估,图 3 和图 4 显示了 MirrorGAN++和之前模型生成的文本到图像合成示例。与 GAN-INT-CLS、StackGAN 和 AttnGAN 相比, STRM-MirrorGAN++方法生成的图像质量更高, 细节更多, 文本图像的语义一致性也更好。这是因为 MCAM 和 STRM 有助于生成具有更多细节和更好语义一致性的细粒度图像。

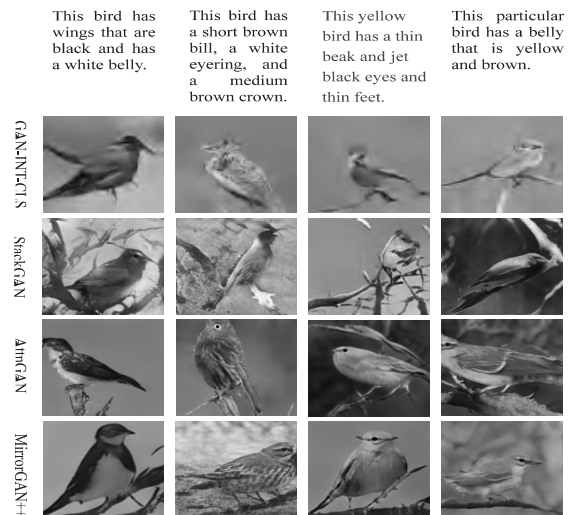


图 3 不同的模型在 CUB 测试集上的文本生成图像

在单主题生成方面,即图 3 中 CUB 数据集上生成的样本, MirrorGAN++模型更好地突出了图像的主题鸟并且细节呈现得更丰富, MirrorGAN++方法能够更好地理解文本描述的逻辑,并呈现出更清晰的图像结构。例如,图 3 的第 1 列、第 2 列和第 4 列中 AttnGAN 生成的示例的鸟的头部到图片的外面,而文中模型就没有这个问题。

在多主题生成方面,即图 4 中的 COCO 数据集生成的样本,当文本描述更复杂且包含多个主题时,生成图像更具挑战性。MirrorGAN++根据最重要的主题精确地捕捉主场景,并合理地安排其余的描述性内容,从而改善了图像的整体结构。例如,在图 4 的第 2 列需要标识浴室所需的组件,而 MirrorGAN++是唯一一种成功的方法。还可以看出,图 4 的第 3 列和第 4 列中 GAN-INT-CLS、StackGAN 和 AttnGAN 生成的示例的形状看起来很奇怪并且缺乏一些细节,而文中模型要

好很多。

视觉结果表明, MirrorGAN++方法生成的图像质量更高, 细节更多, 文本图像的语义一致性也更好。

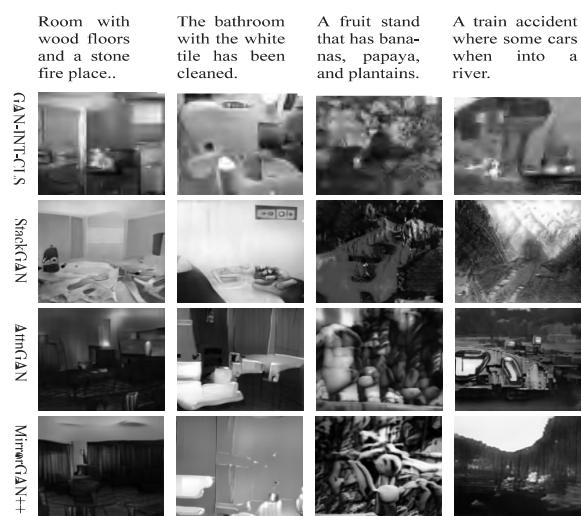


图 4 不同的模型在 COCO 测试集上的文本生成图像

3.3 消融研究

MirrorGAN++关键部分的消融研究: 接下来对所提出的模型进行消融研究。为了验证 STRM 和 MCAM 的有效性, 通过在 MirrorGAN++中移除和添加这些成分进行了几个比较实验。结果见表 3 和表 4。

表 3 不同权重设置下的 MirrorGAN++IS 结果

评价指标	CUB	COCO
MirrorGAN++, no MCAM $\lambda_1 = 0$	3.87 ± 0.07	19.78 ± 0.37
MirrorGAN++, no MCAM $\lambda_1 = 20$	4.71 ± 0.09	30.37 ± 0.11
MirrorGAN++, $\lambda_1 = 5$	4.03 ± 0.06	21.15 ± 0.19
MirrorGAN++, $\lambda_1 = 10$	4.58 ± 0.09	27.58 ± 0.12
MirrorGAN++, $\lambda_1 = 20$	4.82 ± 0.07	31.49 ± 0.13

表 4 不同权重设置下的 MirrorGAN++RI 结果

评价指标	CUB	COCO
MirrorGAN++, no MCAM $\lambda_1 = 0$	38.89 ± 2.37	50.57 ± 1.98
MirrorGAN++, no MCAM $\lambda_1 = 20$	68.32 ± 1.45	87.14 ± 0.97
MirrorGAN++, $\lambda_1 = 5$	45.77 ± 1.27	59.36 ± 1.09
MirrorGAN++, $\lambda_1 = 10$	59.58 ± 0.97	76.32 ± 1.04
MirrorGAN++, $\lambda_1 = 20$	71.43 ± 1.31	89.01 ± 0.94

首先, 参数很重要。在 CUB 数据集上, λ_2 设置为 5, 当 λ_1 从 5 增加到 20 时, IS 从 4.03 增加到 4.82, RI 从 45.77% 增加到 71.43%。在 COCO 数据集上, λ_2 设置为 50, 当 λ_1 从 5 增加到 20 时, IS 从 21.15 增加到 31.49, RI 从 59.36% 增加到 89.01%。将 λ_1 设为 20 作为默认值。

没有 MCAM 和 STRM ($\lambda_1 = 0$) 的 MirrorGAN++ 已经比 StackGAN++^[22] 和 PPGN^[23] 取得了更好的效果。将 STRM 集成到 MirrorGAN++ 中会进一步提高

性能。IS 在 CUB 中从 3.87 分提高到 4.71 分, 在 COCO 中从 19.78 分提高到 30.37 分, RI 呈相同趋势。值得注意的是, 没有 MCAM 的 MirrorGAN++ 已经超过了之前的 AttnGAN (见表 1), 后者也使用了单词级的注意力。这些结果表明, STRM 在帮助生成器获得更好的性能方面很有效。具体来说, STRM 从最后生成的图像重新生成文本描述, 在语义上与给定的文本描述对齐。此外, STRM 与 MCAM 的结合进一步提高了 IS 和 RI, 使最后的结果超越了 MirrorGAN。这些结果表明, 提出的改进的多维度的注意力协同模块 MCAM 和语义文本再生模块 STRM 是非常有效的。

4 结束语

提出了改进的多维度的注意力协同模块 MCAM 和语义文本再生模块 STRM, 以解决具有挑战性的 T2I 生成问题。MCAM 具有从粗到细生成目标图像的级联结构, 利用本地单词注意和全局句子注意逐步增强生成图像的多样性和语义一致性。STRM 通过从生成的图像中重新生成文本描述来进一步监督生成器, 使该图像在语义上与给定的文本描述对齐。通过对两个公共基准数据集的深入实验, 证明了 MirrorGAN++ 优于其他方法。

参考文献:

- [1] 赵树阳, 李建武. 基于生成对抗网络的低秩图像生成方法[J]. 自动化学报, 2018, 44(5): 829-839.
- [2] ZHANG H, XU T, LI H, et al. Stackgan: text to photo-realistic image synthesis with stacked generative adversarial networks[C]//IEEE international conference on computer vision. Venice, Italy: IEEE, 2017: 5908-5916.
- [3] ZHANG Z, XIE Y, YANG L. Photographic text-to-image synthesis with a hierarchically-nested adversarial network[C]//IEEE conference on computer vision and pattern recognition. Salt Lake City, UT, USA: IEEE, 2018: 6199-6208.
- [4] XU T, ZHANG P, HUANG Q, et al. AttnGAN: fine-grained text to image generation with attentional generative adversarial networks[C]//IEEE conference on computer vision and pattern recognition. Salt Lake City, UT, USA: IEEE, 2018: 1316-1324.
- [5] HONG S, YANG D, CHOI J, et al. Inferring semantic layout for hierarchical text-to-image synthesis[C]//IEEE conference on computer vision and pattern recognition. Salt Lake City, UT, USA: IEEE, 2018: 7986-7994.
- [6] 孙 钰, 李林燕, 叶子寒, 等. 多层次结构生成对抗网络的文本生成图像方法[J]. 计算机应用, 2019, 39(11): 3204-3209.
- [7] 陈鑫磊, 陈锻生. 分类重构堆栈生成对抗网络的文本生成图像模型[J]. 华侨大学学报: 自然科学版, 2019, 40(4):

- 549–555.
- [8] QIAO Tingting, ZHANG Jing, XU Duanqing, et al. MirrorGAN: learning text-to-image generation by redescription [C]//IEEE conference on computer vision and pattern recognition. Long Beach, CA, USA; IEEE, 2019: 1505–1514.
 - [9] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets [C]//Advances in neural information processing systems. Montréal, CANADA; NIPS, 2014: 2672–2680.
 - [10] REED S, AKATA Z, YAN X, et al. Generative adversarial text to image synthesis [C]//International conference on machine learning. New York, USA; ICML, 2016: 1060–1069.
 - [11] VINYALS O, TOSHEV A, BENGIO S, et al. Show and tell: a neural image caption generator [C]//IEEE conference on computer vision and pattern recognition. Boston, MA, USA; IEEE, 2015: 3156–3164.
 - [12] KARPATHY A, LI Feifei. Deep visual-semantic alignments for generating image descriptions [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 664–676.
 - [13] ZHU J, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks [C]//IEEE international conference on computer vision. Venice, Italy; IEEE, 2017: 2242–2251.
 - [14] YI Z, ZHANG H, TAN P, et al. Dualgan: unsupervised dual learning for image-to-image translation [C]//IEEE international conference on computer vision. Venice, Italy; IEEE, 2017: 2868–2876.
 - [15] ALMAHAIRI A, RAJESWAR S, SORDONI A, et al. Augmented cyclegan: learning many-to-many mappings from unpaired data [C]//International conference on machine learning. Stockholmsmässan, Stockholm, Sweden; IEEE, 2018: 195–204.
 - [16] ISOLA P, ZHU J, ZHOU T, et al. Image-to-image translation with conditional adversarial networks [C]//IEEE conference on computer vision and pattern recognition. Honolulu, HI, USA; IEEE, 2017: 5967–5976.
 - [17] QIAO T, ZHANG W, ZHANG M, et al. Ancient painting to natural image: a new solution for painting processing [C]//IEEE winter conference on applications of computer vision. Waikoloa Village, HI, USA; IEEE, 2019: 521–530.
 - [18] 陈磊, 李俊. 基于词向量的文本特征选择方法研究 [J]. 小型微型计算机系统, 2018, 39(5): 991–994.
 - [19] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding [C]//Annual conference of the North American Chapter of the association for computational linguistics; human language technologies. New Orleans, Louisiana, USA; NACCL, 2018: 1–16.
 - [20] ANDERSON P, HE X, BUEHLER C, et al. Bottom-up and top-down attention for image captioning and visual question answering [C]//IEEE conference on computer vision and pattern recognition. Salt Lake City, UT, USA; IEEE, 2018: 6077–6086.
 - [21] DENG J, DONG W, SOCHER R, et al. Imagenet: a large-scale hierarchical image database [C]//IEEE conference on computer vision and pattern recognition. Miami, FL, USA; IEEE, 2009: 248–255.
 - [22] HAN Z, TAO X, LI H, et al. StackGAN++: realistic image synthesis with stacked generative adversarial networks [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2019, 41(8): 1947–1962.
 - [23] NGUYEN A, CLUNE J, BENGIO Y, et al. Plug & play generative networks: conditional iterative generation of images in latent space [C]//IEEE conference on computer vision and pattern recognition. Honolulu, HI, USA; IEEE, 2017: 3510–3520.
 - [24] 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.
 - [25] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context [C]//European conference on computer vision. Zurich, Switzerland; Springer International Publishing, 2014: 1–14.
 - [26] SALIMANS T, GOODFELLOW I, ZAREMBA W, et al. Improved techniques for training gans [C]//Advances in neural information processing systems. Barcelona, Spain; NIPS, 2016: 2234–2242.