计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.201100207

基于场景图的段落生成序列图像方法

张玮琪[1,2]　汤轶丰[1,2]　李林燕[3]　胡伏原[1,4]
1 苏州科技大学电子与信息工程学院　江苏 苏州 215009
2 苏州科技大学苏州市大数据与信息服务重点实验室　江苏 苏州 215009
3 苏州经贸职业技术学院　江苏 苏州 215009
4 苏州科技大学苏州市虚拟现实智能交互及应用技术重点实验室　江苏 苏州 215009
(weiqizhang1997@163.com)

摘　要　通过生成对抗网络进行段落生成序列图像的任务已经可以生成质量较高的图像。然而当输入的文本涉及多个对象和关系时,文本序列的上下文信息难以提取,生成图像的对象布局容易产生混乱,生成的对象细节不足。针对该问题,文中在StoryGAN的基础上,提出了一种基于场景图的段落生成序列图像方法。首先,通过图卷积将段落转换为多个场景图,每个场景图包含对应文本的对象和关系信息;然后,预测对象的边界框和分割掩膜来计算生成场景布局;最后,根据场景布局和上下文信息生成更符合对象及其关系的序列图像。在CLEVR-SV和CoDraw-SV数据集上进行测试,该方法可以生成包含多个对象及其关系的64×64像素的序列图像。实验结果表明,在CLEVR-SV数据集上,所提方法的*SSIM*和*FID*比StoryGAN分别提升了1.34%和9.49%;在CoDraw-SV数据集上,所提方法的*ACC*比StoryGAN提高了7.40%。所提方法提高了生成场景的布局合理性,不仅可以生成包含多个对象和关系的图像序列,而且生成的图像质量更高,细节更清晰。

关键词:生成对抗网络;图卷积神经网络;场景布局;文本生成图像
中图法分类号　TP391

Image Stream From Paragraph Method Based on Scene Graph

ZHANG Wei-qi[1,2],TANG Yi-feng[1,2],LI Lin-yan[3] and HU Fu-yuan[1,4]
1 School of Electronic & Information Engineering,Suzhou University of Science and Technology,Suzhou,Jiangsu 215009,China
2 Suzhou Key Laboratory for Big Data and Information Service,Suzhou University of Science and Technology,Suzhou,Jiangsu 215009,China
3 Suzhou Institute of Trade and Commerce,Suzhou,Jiangsu 215009,China
4 Virtual Reality Key Laboratory of Intelligent Interaction and Application Technology of Suzhou,Suzhou University of Science and Technology,Suzhou,Jiangsu 215009,China

Abstract　The task of generating sequence images from paragraphs by generating confrontation networks can already generate higher quality images. However,when the input text involves multiple objects and relationships,the context information of the text sequence is difficult to extract,the object layout of the generated image is prone to confusion,and the generated object details are insufficient. To solve this problem,this paper proposes a method of generating sequence images based on scene graphs based on StoryGAN. First,the paragraph is converted into multiple scene graphs through graph convolution,each scene graph contains the object and relationship information of the corresponding text. Then,the bounding box and segmentation mask of the object are predicted to calculate the scene layout. Finally,according to the scene layout and the context information,a sequence of images more in line with the object and its relationship is generated. Tests on CLEVR-SV and CoDraw-SV data sets show that the method in this paper can generate 64×64-pixel sequence images containing multiple objects and their relationships. Experimental results show that on the CLEVR-SV data set,the *SSIM* and *FID* of this method are improved by 1.34% and 9.49% respectively than StoryGAN. On the CoDraw-SV data set,the *ACC* of this method is 7.40% higher than that of StoryGAN. The proposed method improves the rationality of the layout of the generated scene,not only can generate an image sequence containing multiple

到稿日期:2020-11-30　返修日期:2021-05-26
基金项目:国家自然科学基金(61876121);江苏省重点研发计划项目(BE2017663);江苏省教育厅高等学校自然科学研究面上项目(19KJB520054);江苏省研究生实践创新项目(SJCX20_1119)
This work was supported by the National Natural Science Foundation of China(61876121),Key Research and Development Program of Jiangsu Province(BE2017663),Foundation of Natural Science Research Program in Jiangsu Province Higher Education(19KJB520054) and Graduate Student Practice Innovation Projects in Jiangsu Province(SJCX20_1119).
通信作者:胡伏原(fuyuanhu@mail.usts.edu.cn)

objects and relationships, but also the generated image has higher quality and clearer details.

Keywords Generative adversarial networks, Graph convolutional network, Scene layout, Text-to-image synthesis

1 引言

文本生成图像在近几年受到越来越多的关注，并且一直是计算机视觉中的一个热门研究领域。它已广泛应用于图像处理、文本分析、信息安全、人机交互等领域。随着深度学习的发展，文本生成图像方法取得了显著进展。但由于文本和图像之间结构差异较大、复杂场景建模困难、所需数据集种类单一等因素，它仍被认为是一项具有挑战性的任务。

文本生成图像模型主要分为 VAE(Variational Auto Encoder)模型[1]、DRAW(Deep Recurrent Attention Writer)模型[2]和生成对抗网络(Generative Adversarial Network, GAN)模型[3]。近年来，基于 GAN 的文本生成图像方法取得了显著进步。Reed 等[4]提出了 GAN-INT-CLS 模型，首次利用 GAN 有效地生成以文本描述为条件的 64×64 图像。Xu 等[5]提出的 AttnGAN 模型引入了注意力机制来生成 256×256 图像，实验结果表明，层级条件生成对抗网络(layered conditional GAN)能够自动选择单词级别的条件来生成图像的不同部分。Li 等[6]提出了对象驱动的生成对抗网络(Obj-GANs)，通过关注文本描述中最相关的词和预先生成的语义布局来合成显著对象。Li 等[7]首次提出了段落生成图像序列的任务 StoryGAN，学习从自然语言故事中生成有意义且连贯的图像序列。

最近，基于场景图的生成模型[8]因为可以更明确地表示出复杂句子所传达的信息而受到极大关注。场景图生成是图像和语言强有力的结构化表示，目前已经提出了将句子转换成场景图的方法[9]和通过场景图转换成图像的方法[10-13]。但是当输入的段落包含多个对象和关系的复杂场景时，生成序列图像的连贯性仍是一个具有挑战性的任务，当前从段落生成序列图像的方法与包含多个对象的输入文本相冲突。

针对上述问题，本文在 StoryGAN[7]的基础上，提出了一种基于场景图的段落生成序列图像方法(Scene Graph Generative adversarial networks, SGGAN)。本文通过在生成模型中引入图卷积网络，加强特征之间信息的关联；将生成模型和场景图相结合，准确推理对象间的位置关系，解决了对象布局混乱的问题，生成具有多个对象的序列图像；同时，为了验证本文方法，利用现有的 CLEVR[14]数据集和 CoDraw[15]数据集创建了两个新的数据集，分别称为 CLEVR-SV 和 CoDraw-SV。与现有的算法相比，本文方法能更有效地捕捉故事描述中的对象与对象间的关系。

2 相关工作

2.1 生成对抗网络

生成对抗网络是一种深度学习模型，与传统机器学习方法不同，其最大的特点是引入了对抗机制。对抗的双方由生成器网络(Generator)和鉴别器网络(Discriminator)组成，如图 1 所示。生成器网络学习真实的数据分布 $P_{\text{data}}(x)$，鉴别器网络判断输入数据是来自真实数据还是生成器网络生成的数据。

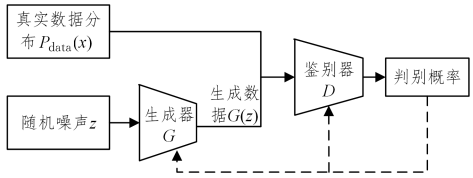


图 1 GAN 模型

Fig. 1 Model of GAN

在训练中，生成器网络 G 输入服从先验分布的随机噪声 z，生成一个类似真实训练样本的数据；鉴别器网络 D 是一个二分类器，用于估计样本来自训练数据而非生成数据的概率，根据输出的概率值区分输入对象是真实图像还是生成图像。生成器网络 G 和鉴别器网络 D 的训练是一个极大极小博弈，定义为：

$$\max_G \min_D V(D,G)=E_{x\sim P_{\text{data}}}[\log D(x)]+E_{x\sim P_z}[\log(1-D(G(z)))] \tag{1}$$

其中，G 表示生成器网络，D 表示鉴别器网络，$E[\cdot]$表示数学期望，V 表示目标函数，x 表示样本，P_{data}表示真实样本的概率分布，P_z表示生成样本的概率分布。

2.2 StoryGAN

StoryGAN 首次提出了故事可视化任务[7]，即给定一个多句子的段落，通过生成一系列图像来可视化故事，每个句子都是一个图像。它更关注动态场景和角色之间的全局一致性，这是任何单个图像或视频生成方法都无法解决的问题。

StoryGAN 包括一个动态跟踪故事流的深度上下文编码器，以及两个故事和图像鉴别器，以提高图像质量和生成序列的一致性。其中上下文编码器又包含 GRU 单元和 Text2Gist 单元两部分。如果每个图像在语义上都能匹配其对应的句子，那么图像序列就是局部一致的。如果所有图像都和可视化的故事 S 一样是全局一致的，那么图像序列就是全局一致的。首先，将段落编码和随机采样；然后编码的向量被提供给基于 RNN 的上下文编码器，以在连续图像生成期间捕获上下文信息，如图 2 所示。

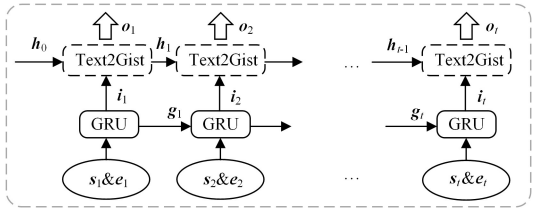


图 2 上下文编码器的框架

Fig. 2 Framework of context encoder

上下文编码器将当前句子和故事编码向量转换成高维特征向量 ***Gist***，用于进一步的图像生成。随着故事的进行，***Gist*** 会动态更新，以反映故事流程中对象和场景的变化。在 Text2Gist 组件中，句子描述被转换成一个过滤器，并根据故事进行调整，这样就可以通过调整过滤器来优化混合过程。类似的想法也被用于动态卷积[16]、注意力模型[5]和元学习[17]。

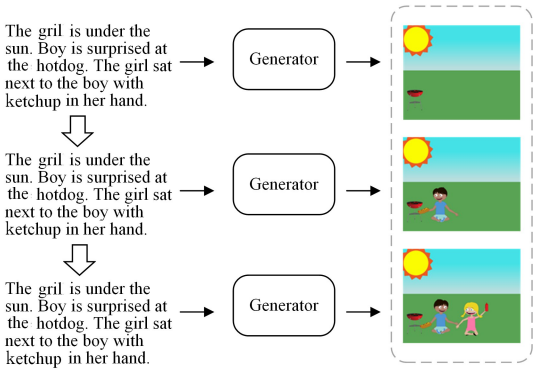
在时间步长 t 时刻，GRU 层输入句子向量$\boldsymbol{s}_t$和高斯噪声 e_t，然后输出向量$\boldsymbol{i}_t$。Text2Gist 单元将 GRU 的输出$\boldsymbol{i}_t$与故事

上下文向量 $\boldsymbol{h}_t$（由故事编码器初始化）相结合，以生成 $\boldsymbol{o}_t$，$\boldsymbol{o}_t$ 对在时间 t 生成图像所需的所有信息进行编码。$\boldsymbol{h}_t$ 由 Text2Gist 单元更新，以反映潜在的上下文信息变化。

3 基于场景图的段落生成序列图像方法

3.1 方法框架

基于场景图的段落生成序列图像的生成对抗网络旨在创建一系列图像来描述输入的段落 S。其流程如图 3 所示。



注：每个句子都用一个图像形象化表示

图 3 输入段落描述生成序列图像

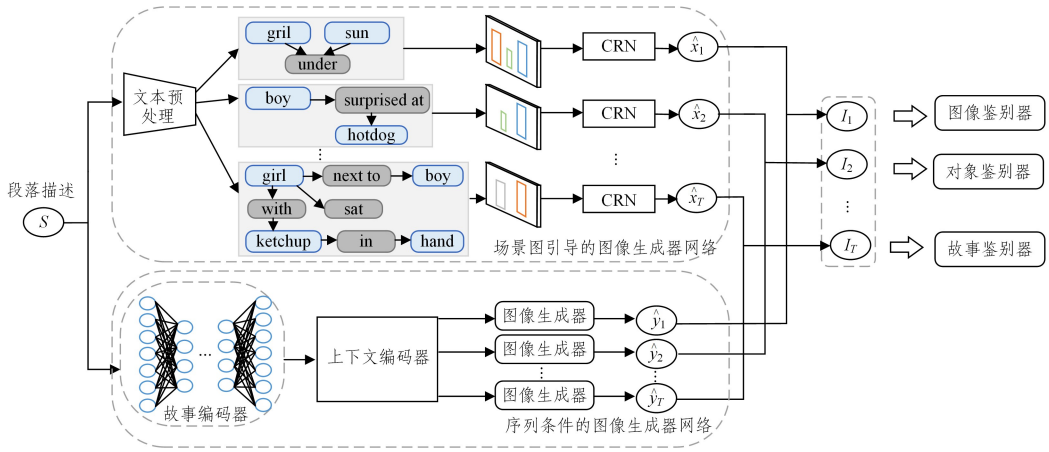
Fig. 3 Input paragraph description to generate sequence images

该网络主要由 3 部分组成：1）场景图引导的图像生成器，它以文本经过处理的场景图为输入，生成与该图相对应的真实图像序列；2）基于序列条件的图像生成器，它把段落 S 编码成一个低维向量 $\boldsymbol{h}_0$，并将输入的句子及其上下文信息编码为每个时间点 t 的向量，引导生成局部和全局一致的图像序列；3）鉴别器（图像鉴别器、对象鉴别器和故事鉴别器），其引导图像生成过程，以确保图像生成的局部一致性和全局一致性。

SGGAN 的整体架构如图 4 所示。给定一个多句子的段落，SGGAN 通过两个生成器分别生成图像。段落 S 由一系列句子 $s=[s_1,s_2,\cdots,s_T]$ 组成，其中长度 T 可以变化。在场景图引导的图像生成器网络中，每个句子通过场景图生成的图像被命名为 $[\hat{x}_1,\hat{x}_2,\cdots,\hat{x}_T]$。在基于序列条件的图像生成模型中，每个句子生成的图像被命名为 $[\hat{y}_1,\hat{y}_2,\cdots,\hat{y}_T]$。两个生成器生成的图像合并后的图像被称为 $[I_1,I_2,\cdots,I_T]$。训练时，真实图像表示为 $P=[p_1,p_2,\cdots,p_T]$。本文通过对 3 个鉴别器网络 D_{img}，D_{obj} 和 D_{str} 进行对抗训练来生成图像，使得生成的图像既逼真又包含可识别的对象，如图 3 所示。

3.2 场景图引导的图像生成器

如图 4 所示，场景图引导的图像生成器包括文本预处理模型、场景图网络、场景布局和级联细化网络。



注：给定一段文字，生成符合对象之间关系且局部和全局一致的图像序列，CRN 为级联细化网络

图 4 SGGAN 的框架

Fig. 4 Framework of SGGAN

本模型使用图像生成网络将场景图转换为图像，段落 S 通过文本预处理模型后，生成每个句子的场景图 G_i。该网络输入场景图 G_i 和噪声 z，并输出图像 $\hat{x}_i=(G,z)$。场景图 G_i 由图卷积网络处理，该网络给出每个对象的嵌入向量，图卷积的每一层沿着图形的边缘混合信息。本文通过使用来自图卷积网络的对象嵌入向量来预测每个对象的边界框和分割掩膜，从而获得来自 G_i 的对象和关系；这些边界框和分割掩膜结合起来形成一个场景布局，它充当了图形和图像域之间的中间点。输出图像 $\hat{x}_i$ 是使用级联细化网络（Cascaded Refinement Network，CRN）[18] 处理场景布局而生成的，CRN 的每个模块都以不断增加的空间比例处理布局，最终生成图像 $\hat{x}_i$。

3.2.1 多个场景图转换

本文通过文本预处理把段落 S 中的每个句子都转换为对应的场景图。从文本生成场景图是一个广泛研究的问题[9]，相当于将图像描述解析为场景图。本文使用先进的方法进行场景图预测，由场景图引导的图像生成器建立在文本预处理模型的输出上。将图像描述解析为场景图的任务定义如下。给定一组对象类 C、一组关系类型 R、一组属性类型 A 和一个句子 s，要将句子 s 解析为场景图 $G=(O,E)$。$O=\{o_1,\cdots,o_n\}$ 是 s 中提到的一组对象，每个 o_i 都是一对 (c_i,A_i)，其中 $c_i\in C$ 是 o_i 的类，$A_i\subseteq A$ 是 o_i 的属性。$E\subseteq O\times R\times O$ 是图中两个对象之间的关系集。例如，给定句子：

$$s=\text{The boy is playing his black football} \tag{2}$$

需要提取两个对象 $o_1=(boy,\varnothing)$ 和 $o_2=(\text{football},\{\text{black}\})$，以及关系

$$\begin{aligned} e_1&=(o_1,playing,o_2)\\ e_2&=(o_1,his,o_2) \end{aligned} \tag{3}$$

集合 C,R 和 A 由训练数据中存在的所有类和类型组成。最终得到符合文本描述的多个场景图。

3.2.2 场景布局的生成

以端到端的方式处理场景图,需要一个能够对图进行本地操作的神经网络模块。为此,本文使用由几个图形卷积层组成的图卷积网络。

传统的 2D 卷积层将特征向量的空间网格作为输入,并产生新的向量空间网格作为输出,其中每个输出向量都是其相应输入向量的局部邻域的函数,这样,卷积便聚集了输入的局部邻域的信息。单个卷积层可以通过使用输入中所有邻域的权值共享对任意形状的输入进行操作。

本文的图形卷积层执行类似的功能:给定一个场景图,每个节点和边都有维数 D_{in} 的向量,它为每个节点和边计算新的维数 D_{out} 的向量。输出向量是其相应输入向量的邻域的函数,因此每个图卷积层沿着图形的边缘传播信息。图卷积层对图形的所有边应用相同的函数,允许单个层对任意形状的图形进行操作。本文用一系列图形卷积层处理输入场景图,给出每个对象的嵌入向量,该向量聚合图形中所有对象和关系的信息。

为了生成图像,模型必须从图形域移动到图像域,因此本文使用对象嵌入向量来计算场景布局。该场景布局给出了要生成的图像的粗略 2D 结构,通过使用对象布局网络预测每个对象的分割掩膜和边界框来计算场景布局。

对象布局网络输入一个 D 维的嵌入向量 $\boldsymbol{v}_i$,将其传递给掩膜回归网络和边界框回归网络,分别预测 $M\times M$ 的权重矩阵和边界框 $\hat{b}_i=(x_0,y_0,x_1,y_1)$。掩膜回归网络包含了在非线性 sigmoid 函数上的几个转置卷积,故生成掩膜的值在(0,1)范围内,边界框回归网络是 MLP(Multi-Layer Perceptron)。

将嵌入向量 $\boldsymbol{v}_i$ 与掩膜 $\hat{m}_i$ 相乘,得到一个 $D\times M\times M$ 形状的掩膜嵌入,然后使用双线性插值将其放到边界框的位置,生成对象布局。计算所有对象布局的和即可得到场景布局。在训练期间使用 Ground Truth 的边界框 b_i 来计算场景布局;在测试时改用预测的边界框 $\hat{b}_i$ 来计算场景布局。

3.2.3 CRN(级联细化网络)生成序列图像

给定场景布局,必须合成一个符合布局中给定的对象位置的图像。在这个任务中,本文使用了级联细化网络(CRN)[18]。如图 5 所示,CRN 由一系列卷积细化模块组成,模块之间的空间分辨率加倍,这允许 CRN 以从粗到细的方式生成图像。

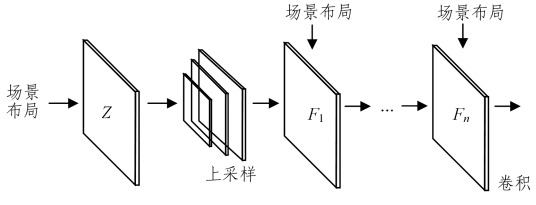


图 5 级联细化网络的框架

Fig. 5 Framework of cascaded refinement network

每个模块接收场景布局和前一个模块的输出 F_i 作为输入。这些输入通过通道连接,并传递给一个 3×3 卷积层,然后在传递到下一个模块之前,使用最近邻插值对输出进行上采样。第一个模块接收高斯噪声 $z\sim p_z$ 作为输入,最后一个模块的输出被传递到两个最终卷积层以产生输出图像 $[\hat{x}_1,\hat{x}_2,\cdots,\hat{x}_T]$。

3.3 基于序列条件的图像生成器网络

整个基于序列条件的图像生成器网络可以看作是一个序列化的条件生成对抗网络框架,该网络包含故事编码器(Story Encoder)、上下文编码器(Context Encoder)和图像生成器(Image Generator)。

在该网络中,段落 $\boldsymbol{S}$ 的每个句子在输入到故事编码器之前都已经使用预先训练的句子编码器编码成嵌入向量,每个句子都是通过向量 $\boldsymbol{s}_t\in\mathbb{R}^{128}$ 编码的。本文假设 $\boldsymbol{s}_t$ 和 $\boldsymbol{S}$ 都是编码的特征向量,而不是原始文本。

3.3.1 故事编码器

故事编码器在图 4 的序列条件图像生成器中给出。根据 StackGAN[19] 中的调节机制,故事编码器 $E(\cdot)$ 学习从故事 $\boldsymbol{S}$ 到低维嵌入向量 $\boldsymbol{h}_0$ 的随机映射。它通过使用随机采样处理原始故事空间中特征空间不连续的问题。为了避免过拟合和生成模式坍塌到单个生成点上,引入了正则化项,即计算生成分布和标准正态分布之间的 KL(Kullback-Leibler)散度[19]:

$$\mathscr{L}_{\text{KL}}=KL(\mathcal{N}(\mu(\boldsymbol{S}),\text{diag}(\sigma^2(\boldsymbol{S})))\,\|\,\mathcal{N}(0,I)) \tag{4}$$

3.3.2 上下文编码器

视频生成与故事可视化密切相关,它通常假设静态背景具有平滑的运动过渡,需要静态和动态特征的不相交嵌入[20-22]。在本模块的图像生成任务中,两个挑战的不同之处在于角色、动作经常在图像之间变化,如图 3 所示。这需要解决两个问题:1)如何更新背景信息以有效捕捉背景变化;2)当生成每幅图像时,如何组合新的输入和随机噪声,以可视化可能发生显著变化的字符变化。

本文运用了一个基于深度 RNN 的上下文编码器来解决上述问题,以在连续图像生成期间捕获上下文信息。上下文可以定义为故事中对当下生成有用的任何相关信息。经过图像生成器可得到该网络的序列图像 $[\hat{y}_1,\hat{y}_2,\cdots,\hat{y}_T]$。

两个生成器网络生成的图像融合后最终生成的结果如下:

$$I_i=Conv(Concat(\hat{x}_i,\hat{y}_i)) \tag{5}$$

其中,$Concat(\cdot)$ 和 $Conv(\cdot)$ 分别表示信道级联和卷积运算。

3.4 鉴别器

SGGAN 使用 3 个鉴别器:图像鉴别器 D_{img}、对象鉴别器 D_{obj} 和故事鉴别器 D_{str},用于确保图像生成的局部一致性和全局一致性。

给定以 h_0 编码的初始上下文信息,图像鉴别器判别生成的图像 I_t 是否与句子 $\boldsymbol{s}_t$ 匹配。它通过将生成的三元组 $\{\boldsymbol{s}_t,\boldsymbol{h}_0,I_t\}$ 与真实的三元组 $\{\boldsymbol{s}_t,\boldsymbol{h}_0,I_t\}$ 进行比较来实现。与以往的文本生成图像方法不同,图像生成器网络根据上下文和相同的

句子可以生成显著不同的图像，因此将编码的上下文信息提供给鉴别器是很重要的。例如，输入文本“绿色橡胶球体位于左下角。然后在它的上边添加一个黄色金属的立方体。”第二个图像在没有上下文（即第一句话）的情况下会有很大不同。

对象鉴别器D_{obj}确保图像中的每个对象看起来都是真实的，它的输入是对象的像素，使用双线性插值法将对象裁剪并重新缩放到固定大小。D_{obj}除了将每个物体分类为真或假之外，还使用辅助分类器[23]（auxiliary classifier）来预测物体的类别，从而确保每个物体都是可识别的。对象鉴别器D_{obj}和场景图引导的图像生成器网络都试图将D_{obj}正确分类对象的概率最大化。

在故事鉴别器中，段落里图像和句子的特征向量被连接起来。图像和文本特征的乘积被输入到具有 sigmoid 非线性的完全连接层，以预测生成结果是假的还是真实的故事。

故事鉴别器有助于增强给定故事的生成图像序列的全局一致性。故事鉴别器的整体架构如图 6 所示。其左侧是图像编码器，它将图像序列编码成特征向量$\boldsymbol{E}_{img}(I)=[\boldsymbol{E}_{img}(I_1),\cdots,\boldsymbol{E}_{img}(I_t)]$的序列，其中 I 是真实的或生成的图像。这些向量被连接成单个向量，如图 6 中的蓝色矩形所示。图 6 的右侧是文本编码器，它将段落 S 编码成一系列特征向量：

$$\boldsymbol{E}_{txt}(S)=[\boldsymbol{E}_{txt}(s_1),\cdots,\boldsymbol{E}_{txt}(s_T)] \tag{6}$$

这些向量被连接成一个大向量，如图 6 中的灰色矩形所示。图像编码器是一个深度卷积网络，文本编码器是一个多层感知器。两者输出相同的维度向量。

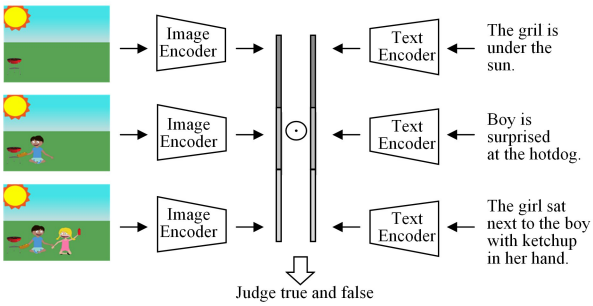


图 6 故事鉴别器D_{str}（电子版为彩色）

Fig. 6 Framework of story discriminator D_{str}

全局一致性分数的计算如下：

$$D_{str}=\sigma(\omega^{T}\boldsymbol{E}_{img}(I)\odot\boldsymbol{E}_{txt}(\boldsymbol{S}))+b \tag{7}$$

其中，$\odot$表示对应元素相乘。权重 ω 和偏置 b 的值在输出层学习得到。σ 是一个 sigmoid 函数，它将分数归一化为[0，1]的值。通过将每个句子和图像配对，故事鉴别器可以同时考虑局部匹配和全局一致性。图像和故事鉴别器都是在正负对上训练的。

4 实验结果及分析

4.1 实验环境和数据集

本文算法采用深度学习框架 PyTorch，实验环境为 ubantu14.04 操作系统，使用 4 块 NVIDIA 1080Ti 图形处理器（GPU）加速运算，最终得到训练完成的模型。本文在数据集 CLEVR-SV 和 CoDraw-SV 上训练模型并生成 64×64 的图像。

CLEVR（Compositional Language and Elementary Visual Reasoning）数据集[14]是最初用于研究 VQA 系统执行视觉推理能力的诊断数据集，称为组合式语言和初级视觉推理诊断数据集。CLEVR 包含 10 万张经过渲染的图像和大约 100 万个自动生成的问题，其中有 85.3 万个问题是互不相同的。本文通过从随机分配的对象布局中生成图像，为生成序列的图像修改了这些数据（如图 7 顶行的例子，每 4 张图像组成一组序列，每张图依次添加一个图形）。在本文中将该数据集命名为 CLEVR-SV，以区别于现有的 CLEVR 数据集。

CoDraw[15]数据集是基于抽象场景构成的数据集。通过 Zitnick 等[24]的方法，将原始数据生成了 1 002 组抽象场景和描述，每组包含语义相似的 10 个场景，大多数场景包含 6 个对象（平均值 6.67）。CoDraw 数据集收集了由人类玩家之间交换的 138 000 条信息组成的 10 000 个对话。本文修改了 CoDraw 数据集，以 3 个连续的图像形成一个故事，以使其适应段落生成图像序列的任务。最后，本文以 14 536 个描述故事对作为数据集，其中 12 000 对用于训练，剩下的 2 536 对用于测试。本文称该数据集为 CoDraw-SV，以区别于最初的 CoDraw 数据集。

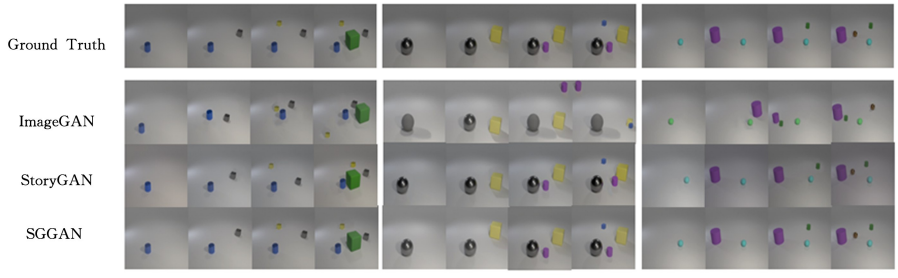


图 7 不同方法在 CLEVR-SV 数据集上的比较（电子版为彩色）

Fig. 7 Comparison of different methods on CLEVR-SV dataset

4.2 评价标准

4.2.1 CLEVR-SV 数据集评价标准

本文在 CLEVR-SV 数据集中主要使用 SSIM（Structural Similarity Index Measure）和 FID（Frechet Inception Distance）两种评价标准，对生成的图像与 Ground Truth 之间的质量和相似性进行评价。

(1)SSIM[25]指标

SSIM是一种衡量两幅图片相似度的指标，用于得到两幅图像的相似性，进而确定生成的图像是否与输入的文本相符。其定义为：

$$SSIM(x,y)=[l(x,y)]^{\alpha}[c(x,y)]^{\beta}[s(x,y)]^{\gamma} \tag{8}$$

其中，$l(x,y)$是亮度比较，定义为：

$$l(x,y)=\frac{2\mu_x\mu_y+c_1}{\mu_x^2+\mu_y^2+c_1} \tag{9}$$

$c(x,y)$是对比度比较，定义为：

$$c(x,y)=\frac{2\sigma_{xy}+c_2}{\sigma_x^2+\sigma_y^2+c_2} \tag{10}$$

$s(x,y)$是结构比较，定义为：

$$s(x,y)=\frac{\sigma_{xy}+c_3}{\sigma_x\sigma_y+c_3} \tag{11}$$

其中，μ_x和μ_y分别代表x，y的平均值，σ_x和σ_y分别代表x，y的标准差，σ_{xy}代表x，y的协方差，c_1，c_2，c_3为常数，避免分母为0带来的系统错误。

(2)FID[26]指标

FID是用来计算真实图像与生成图像的特征向量间距离的一种度量，这里的特征向量是由Inception v3 Network得到的。FID定义为：

$$FID=\|\mu_r-\mu_g\|^2+T_r(\boldsymbol{\Sigma}_r+\boldsymbol{\Sigma}_g-2(\boldsymbol{\Sigma}_r\boldsymbol{\Sigma}_g)^{1/2}) \tag{12}$$

其中，μ_r代表真实图片的特征的均值，μ_g代表生成图片的特征均值，$\boldsymbol{\Sigma}_r$代表真实图片的特征的协方差矩阵，$\boldsymbol{\Sigma}_g$代表生成图片的特征的协方差矩阵，T_r代表矩阵的迹。

4.2.2 CoDraw-SV数据集评价标准

在CoDraw-SV数据集中主要使用定量评估和定性评估对生成图像的质量和语义一致性进行评价。

(1)Accuracy(ACC)指标

ACC进行定量评估，用来预测正确字符数量占总数量的比例。为了比较分类准确率，选择8个字符："boy""girl""tree""cat""dog""sun""sandbox""table"。

(2)Human Rank

Human Rank用于定性评估。在测试集中随机选择30组文本描述，对于每个段落，生成模型生成3个图像序列。将3个图像序列和对应的文本描述分给不同的人按不同的方法进行图像质量的排名，最后计算平均排名来评价生成图像的文本匹配度和角色一致性。

在先前的工作中，StoryGAN已提出段落生成序列图像[7]的任务。本文对比了修改后的ImageGAN和StoryGAN，旨在展示本文方法生成的复杂场景的图像符合输入场景图的对象及其关系。ImageGAN遵循文献[4,26]的工作，不使用故事鉴别器、故事编码器和上下文编码器，每个图像都是独立生成的。然而，为了进行合理的比较，本文将s_t、编码的故事S和一个噪声项连接起来作为输入，否则，模型无法完成任务。与ImageGAN相比，StoryGAN包含了额外的上下文编码器和故事鉴别器[7]，缺少本文的场景图引导的图像生成器网络。

4.3 实验结果

4.3.1 CLEVR-SV数据集实验结果

表1给出了CLEVR-SV测试集中SGGAN，Image-GAN和StoryGAN模型的*SSIM*和*FID*评分。与ImageGAN相比，SGGAN的*SSIM*提高了15.03%，*FID*降低了34.76%；与StoryGAN相比，SGGAN的*SSIM*提高了1.34%，*FID*降低了9.49%。

表1 CLEVR-SV中不同方法生成64×64图像的评分

Table 1 Scoring of 64×64 images generated by different methods in CLEVR-SV

模型	*SSIM*	*FID*
ImageGAN	0.592	51.2
StoryGAN	0.672	36.9
SGGAN	**0.681**	**33.4**

图7给出了不同方法在CLEVR-SV数据集上的比较结果。文本输入信息是当前对象的属性和相对位置，由两个表示其坐标的实数给出。例如，图7左栏的第一幅图像是从描述"blue, small, metal, cylinder, (−2.3, 2.6)"中生成的。所有对象的描述都以相同的方式。给定描述，生成的对象的外观应该与Ground Truth相差很小，并且它们的相对位置应该相似。

通过对比可以看出，ImageGAN[4]无法保持输入文本的一致性，当对象数量增加时，会混淆属性，如第二组序列图像中生成了错误的橡胶属性的球体。StoryGAN可以通过故事编码器来解决段落一致性的问题，然而在判断对象之间的关系和对象的相对位置时，图像无法正确生成。与ImageGAN和StoryGAN模型相比，本文SGGAN模型生成的图像更加平滑，与参考图像的差异更小，对象之间的关系更准确。图8为图7局部放大后文本生成图像的不同方法对比。可以看出，与Ground Truth相比，ImageGAN难以准确生成连贯图像，银色球体由输入条件的"金属"变为了"橡胶"，紫色圆柱体与蓝色正方体的位置相差过大，黄色正方体大小不准确；StoryGAN中能够生成较连贯的图像，但当生成多个对象时，可以看出紫色圆柱体更靠近右侧，对象布局较混乱。这表明在这项任务中，本文方法明显优于其他方法，本文模型生成的序列图像质量更高。与使用文本描述作为输入的方法相比，使用场景图作为输入的方法更有利于生成包含多个对象和关系的复杂图像。

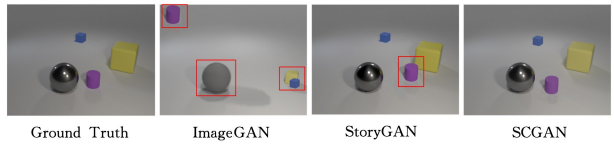


图8 不同方法差异的比较(电子版为彩色)

Fig. 8 Comparison of differences among different methods

4.3.2 CoDraw-SV数据上的实验结果

表2列出了CoDraw-SV测试集中SGGAN，ImageGAN和StoryGAN模型的准确性和Human Rank评分。其中Upper Bound是测试集中真实图像的分类准确性。可以看出，本文

方法与 ImageGAN 相比，准确性提高了31.81%；与 StoryGAN 相比，准确性提高了 7.40%。

表 2 CoDraw-SV 数据集中不同方法生成 64×64 图像的评分

Table 2 Scoring of 64×64 images generated by different methods in CoDraw-SV

模型	准确率	Human Rank
Upper Bound	0.88	—
ImageGAN	0.22	1.87±0.03
StoryGAN	0.27	1.13±0.03
SGGAN	**0.29**	**1.08±0.02**

图 9 给出了段落生成序列图像模型的样本图像，输入文本在左侧给出。训练一个新的文本编码器几乎不会带来性能提升，因此上下文编码器使用带有固定预训练参数的通用编码。由图 9 可以看出，使用场景图作为输入的方法更有利于生成包含多个对象和关系的复杂图像。ImageGAN 无法生成一致的图像序列，且角色的外观在图像序列中是不一致的。相比之下，SGGAN 生成的图像质量更高，在具有多个对象及关系的图像生成中能更好地把握其关系。

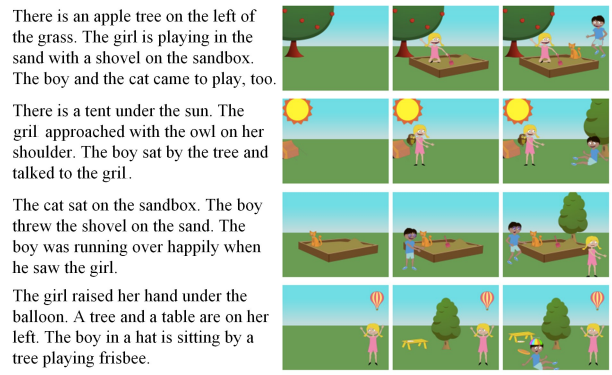


图 9 SGGAN 在 CoDraw-SV 数据集上的生成结果

Fig. 9 SGGAN generated results in CoDraw-SV dataset

结束语 针对现有的段落到序列图像生成模型无法准确生成包含多个对象和关系的图像的问题，本文提出了一种基于场景图的段落生成序列图像模型，它可以基于多个对象生成具有一致性的序列图像。该模型利用图卷积网络对场景图进行特征提取，并引入级联细化网络，不仅提高了生成模型的信息预测能力，还增强了序列图像的连贯性。最终的定量和定性实验结果表明，SGGAN 方法在 SSIM，FID 和 ACC 评价指标中均表现优异，本文模型生成的序列图像中对象之间的关系更符合事实，图像质量更高。但本文依赖大量图像标注信息的数据集，对于训练中未出现的对象及其关系的生成效果较差。未来我们将进一步探索在无标签数据集上生成图像的方法。

参 考 文 献

[1] KINGMA D P，WELLING M. Auto-encoding variational bayes[C]// Proceedings of the International Conference on Learning Representations. 2014.

[2] BA J，MNIH V，KAVUKCUOGLU K. Multiple object recognition with visual attention[C]// International Conference on Learning Representations. 2015.

[3] GOODFELLOW I，POUGET-ABADIE J，MIRZA M，et al. Generative adversarial nets[C]// Advances in Neural Information Processing Systems. 2014：2672-2680.

[4] REED S，AKATA Z，YAN X，et al. Generative adversarial text to image synthesis[C]// Proceedings of the 33rd International Conference on Machine Learning. 2016.

[5] XU T，ZHANG P，HUANG Q，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. 2018：1316-1324.

[6] 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. 2019：12174-12182.

[7] LI Y，GAN Z，SHEN Y，et al. Storygan：A sequential conditional gan for story visualization[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019：6329-6338.

[8] JOHNSON J，GUPTA A，FEI-FEI L. Image generation from scene graphs[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018：1219-1228.

[9] XU D，ZHU Y，CHOY C B，et al. Scene graph generation by iterative message passing[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017：5410-5419.

[10] YANG X，TANG K，ZHANG H，et al. Auto-encoding scene graphs for image captioning[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019：10685-10694.

[11] DHAMO H，FARSHAD A，LAINA I，et al. Semantic Image Manipulation Using Scene Graphs[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020：5213-5222.

[12] SHI J，ZHANG H，LI J. Explainable and explicit visual reasoning over scene graphs[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019：8376-8384.

[13] LAN H，LIU Q Y. Image generation from scene graph with graph attention network[J]. Journal of Image and Graphics，2020，25(8)：1591-1603.

[14] JOHNSON J，HARIHARAN B，Van Der MAATEN L，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.

[15] JIN-HWA K，NIKITA K，XINLEI C，et al. Codraw：Collaborative drawing as a testbed for grounded goal-driven communication[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019：6495-6513.

[16] CHEN Y，DAI X，LIU M，et al. Dynamic convolution：Attention over convolution kernels[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020：11030-11039.

[17] LIU H, SOCHER R, XIONG C. Taming maml: Efficient unbiased meta-reinforcement learning[C]// International Conference on Machine Learning. 2019: 4061-4071.

[18] CHEN Q, KOLTUN V. Photographic image synthesis with cascaded refinement networks[C]// Proceedings of the IEEE International Conference on Computer Vision. 2017: 1511-1520.

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

[20] FU T J, WANG X, GRAFTON S, et al. Iterative Language-Based Image Editing via Self-Supervised Counterfactual Reasoning[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. 2020: 4413-4422.

[21] CHAN C, GINOSAR S, ZHOU T, et al. Everybody dance now [C]// Proceedings of the IEEE International Conference on Computer Vision. 2019: 5933-5942.

[22] TULYAKOV S, LIU M Y, YANG X, et al. Mocogan: Decomposing motion and content for video generation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 1526-1535.

[23] SHEN Y, GU J, TANG X, et al. Interpreting the latent space of gans for semantic face editing[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 9243-9252.

[24] ZITNICK C L, PARIKH D. Bringing semantics into focus using visual abstraction[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013: 3009-3016.

[25] SARA U, AKTER M, UDDIN M S. Image quality assessment through FSIM, SSIM, MSE and PSNR — a comparative study [J]. Journal of Computer and Communications, 2019, 7(3): 8-18.

[26] HEUSEL M, RAMSAUER H, UNTERTHINER T, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium[C]// Advances in Neural Information Processing Systems. 2017: 6626-6637.

ZHANG Wei-qi, born in 1997, postgraduate. Her main research interests include deep learning and computer vision.

HU Fu-yuan, born in 1978, Ph.D, professor, Ph.D supervisor, is a member of China Computer Federation. His main research interests include machine learning and computer vision.

（责任编辑：柯颖）