Image Generation from Scene Graphs

朱婕 21821291
浙江大学
浙江杭州
2057433989@qq.com

摘要

本文主要是介绍和评论*CVPR 2018*中的论文《*Image Generation from Scene Graphs*》。为了准确地从包含多个物体及其之间关系的复杂句子中生成图像，李飞飞等人提出了先用场景图表示文本描述，再从场景图中生成图像的思想，如此能够对对象及其关系进行明确的推理。这篇论文提出的图像生成网络模型用图卷积处理输入场景图，然后根据对象的*bounding boxs*和*segmentation masks*等计算场景布局，并且用级联细化网络将布局转换成图像。这个级联细化网络对一对判别器进行对抗训练，以确保输出的真实性。该模型直接采用事先构建好的场景图作为输入，生成对应的图像，是一个端到端的模型。其中，场景布局的计算是整个图像生成网络的瓶颈，越贴近场景图的场景布局，越有利于生成准确相符的图像。

1. 相关背景和特色

现有的由文本到图像生成的方法主要是由递归神经网络（recurrent neural networks，RNN）和生成对抗网络（Generative Adversarial Networks，GAN）来实现的。其中由代表性的有stackGAN [1]，它在生成花鸟方面的效果确实很棒，并且能达到256*256的高分分辨率。在这篇文章之前，生成图像的分辨率几乎都局限在64*64。stackGAN生成图像分为两个阶段：1)：根据给定描述画出物体大致轮廓和基本的颜色信息，生成低分辨率的图像；2)：把阶段1)的结果和文本描述作为输入，生成带有逼真细节的高分辨率图片。

尽管这些方法生成的效果令人惊艳，但是当它们碰到语言描述中包含多个物体及物体间存在复杂的关系时，往往束手无策。

复杂句子所传达的信息通常可以更明确地表示为对象及其关系的场景图。场景图是用有向图表示场景，其中节点表示对象，边表示对象之间的关系。它们已被用于语义图像检索，评价和改进图像标注；还出现了将句子转换为场景图和从图像预测场景图的方法。作者的团队先将文字描述转换为场景图 [2]，然后由场景图作为模型的输入，通过对场景图设置条件，使模型能够显示地推理对象及其关系，最后得到包含多个对象和关系的复杂图像。

2. 思路和方法

下面首先介绍了作者所遇到的困难和解决思路，然后简单讲解了图像生成模型的结构，最后说明了模型是如何训练的以及专门为该模型设计的六个损失函数。

2.1. 难点和解决方法

该论文提出的模型输入场景图，输出与之对应的的真实图片。构建一个这样的模型主要有一下三个方面的挑战：

1) 需要处理结构化图片输入的方法。
2) 生成的每一个物体都必须看起来真实，而且能正确反应多个物体的空间透视关系。
3) 确保生成图片看起来是自然和谐不别扭的。

作者首先使用图卷积网络（graph convolution network，GCN）处理场景图，沿着图中的边传递信息，然后通过预测图中所有对象的bounding boxs和segmentation

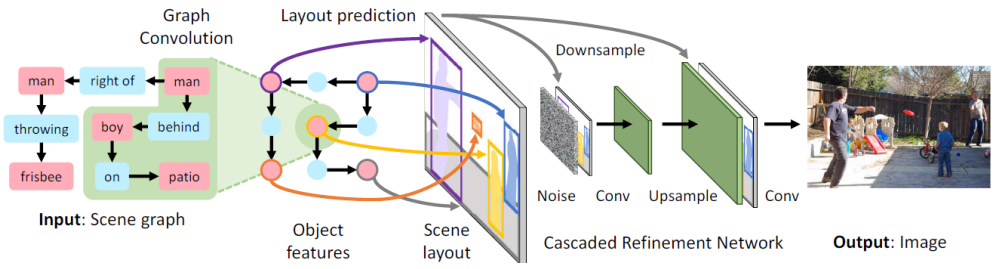

图 1. 图像生成网络概览

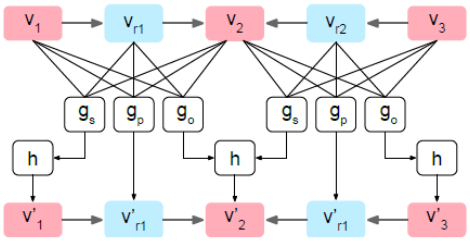

图 2. 单个图卷积网络层实例

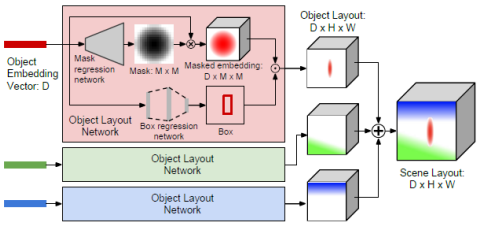

图 3. 场景布局的计算

masks来构建场景布局（scene layout），从而建立从象征型结构化的图像输入到二维的图像输出之间的联系。最后，用级联细化网络（cascaded refinement network，CRN），以递增的空间尺度处理布局，以生成符合预测布局的图像。为了是的生成的图像是真实自然的并且包含可识别的对象，作者将生成网络与一对判别网络一起进行对抗寻训练。这对判别器分别针对图像中的对象和整个图像，从而解决了上述的后两个难点。

2.2. 图像生成模型

模型的结构如图 1所示，输入描述对象及其之间关系的场景图，场景图被GCN处理，GCN计算每个对象嵌入向量。这些向量被传入到Object Layout Network中用于预测对象的bounding boxes和Segmentation masks，将bounding boxes和Segmentation masks结合就可以得到对象的布局，将所有对象布局结合就能形成场景布局（scene layout）。然后用级联细化网络CRN将布局转换为图像。模型训练的时候模型观察真实的bounding boxes和Segmentation masks，测试的时候这些用的都是预测值。

GCN由几个图卷积层构成。类似于传统的二维卷积层以特征向量的空间网格为输入，生成新的空间向量网格作为输出，GCN接收节点和边都是D_{in}维向量的图像作为输入，为每个节点和边计算新的D_{out}维向量。每个输出向量都是其对应输入向量的局部邻域的函数；通过这种方式，卷积将输入的局部邻域的信息沿着边聚集起来。一个图卷积层可以对任意形状的输入进行操作，通过在所有图的边中使用同一个函数。

图卷积层的操作具体来说，就是给定任意对象$o_i \in O$的输入向量$v_i \in \mathbb{R}^{D_{in}}$和边$(o_i, r, o_j) \in E$的输入向量$v_r \in \mathbb{R}^{D_{in}}$，使用函数$o_1, o_2$和$o_3$计算它们的输出向量$v_i', v_r' \in \mathbb{R}^{D_{out}}$,如图 2所示。图 2展示了一个图卷积层的计算过程，假设图中有3个节点o_1, o_2和o_3，还有两条边(o_1, r_1, o_2)和(o_3, r_2, o_2)。沿着边，三个输入向量被传入函数g_s, g_p和g_o；其中g_p直接计算边的输出向量$v_r' = g_p(v_i, v_r, v_j)$，而$g_s$和$g_o$沿着边计算对象的候选输出向量集合

$$V_i^s = \{g_s(v_i, v_r, v_j) : (o_i, r, o_j) \in E\} \quad (1)$$

$$V_i^o = \{g_o(v_j, v_r, v_i) : (o_j, r, o_i) \in E\} \quad (2)$$

其中，V_i^s表示沿着表示对象o_i的节点的出边计算的候选向量的集合，即沿着出边聚集周围的信息，而V_i^O则是沿着节点的出边聚集信息。之后，这些候选向量被传入一个对称的池化函数h以计算一个对象的输出向量，对象o_i的输出为$v_i' = h(V_i^s \cup V_i^o)$。

作者利用一个堆叠的图卷积层对输入场景图进行处理，给出了每个对象的嵌入向量，该向量集合了图中所有对象和关系的信息。为了生成图像，必须从图域移动到图像域。为此，作者利用如图 3所示的对象布局网络，用对象嵌入向量计算场景布局，给出了生成图像的粗二维结构。每个对象的嵌入向量传递给一个对象布局网络以预测其布局，将所有对象的布局加起来就得到了场景布局。对象布局网络由两部分组成，一部分时Mask回归网络，一部分是Box回归网络，分别预测对象的软二值分割掩码和包围框；将这两部分的输出与嵌入向量结合使用双线性插值产生的对象布局。场景布局是所有对象布局之和。

CRN由一系列的卷积细化模块构成，每个模块之间是2倍空间分辨率的关系，这就允许一种由粗到精的方式去生成图片。每个模块结构根据模块输入的分辨率下采样（缩小）后的场景布局和前一个模块的输出，这些输入串联并传递到一对3*3的卷积层，然后在传递到下一个模块之前对输出使用最近邻插值法进行上采样（放大）。

2.3. 训练

图像生成模型是和一对判别器（图像判别器D_{img}和对象判别器D_{obj}）一起训练的。一般而言，一个判别器D通过最小化如下的损失函数，尝试尽可能正确的判断一个输入x的真假：

$$\mathcal{L}_{GAN} = \mathop{\mathbb{E}}_{x\sim p_{\text{real}}} \log D(x) + \mathop{\mathbb{E}}_{x\sim p_{\text{fale}}} \log(1 - D(x)) \quad (3)$$

其中$x \sim p_{\text{fake}}$是生成网络的输出，生成网络尽可能干扰判别器的判断以最小化$\mathcal{L}_{GAN}$，即尽量生成看起来真实的图像。

图像判别器D_{img}确保生成的图像的整体外观是真实的，它将规则间隔，重叠的图像块集合分类为真实或者伪造。

对象判别器D_{obj}确保图像中的每个对象看起都是真实的，其输入是一个对象的像素，使用双线性插值法裁剪并重新缩放到固定大小。除了将每个对象分类为真实的还是假的，对象判别器还确保每个对象都可以使用预测对象类别的辅助分类器来识别；对象判别器和生成网络都企图最大化对象判别器正确分类对象的概率。

联合训练生成网络和两个判别器，生成网络被训练以最小化以下六个损失的加权和：

- Box loss $\mathcal{L}_{box} = \sum_{i=1}^{n} \left\| b_i - \hat{b}_i \right\|_1$：惩罚真实box和预测box的差距
- Mask loss $\mathcal{L}_{mask}$：用位级交叉熵惩罚真实mask和预测mask之间的差距
- Pixel loss $\mathcal{L}_{pix} = \|I - \hat{I}\|_1$：惩罚真实图像和生成图像的差距
- Image adversarial loss $\mathcal{L}_{GAN}^{img}$：针对图像判别器，使得生成的图像真实自然
- Object adversarial loss $\mathcal{L}_{GAN}^{Obj}$：针对对象判别器，使得每一个生成的对象是真实的
- Auxiliarly classifier loss $\mathcal{L}_{AC}^{obj}$：还是针对对象判别器，确保对象判别器能将每一个产生的对象正确分类

Method	Inception	
	COCO	VG
Real Images (64×64)	16.3 ± 0.4	13.9 ± 0.5
Ours (No gconv)	4.6 ± 0.1	4.2 ± 0.1
Ours (No relationships)	3.7 ± 0.1	4.9 ± 0.1
Ours (No discriminators)	4.8 ± 0.1	3.6 ± 0.1
Ours (No D_{obj})	5.6 ± 0.1	5.0 ± 0.2
Ours (No D_{img})	5.6 ± 0.1	$\mathbf{5.7 \pm 0.3}$
Ours (Full model)	$\mathbf{6.7 \pm 0.1}$	5.5 ± 0.1
Ours (GT Layout, no gconv)	7.0 ± 0.2	6.0 ± 0.2
Ours (GT Layout)	$\mathbf{7.3 \pm 0.1}$	$\mathbf{6.3 \pm 0.2}$
StackGAN [59] (64×64)	$\mathbf{8.4 \pm 0.2}$	-

图 4. 模型简化测试结果

3. 相关实验

作者进行了一系列实验证明该模型具有以下特点：

1) 可以生成多对象的场景，甚至是同一类型对象生成多种实例
2) 生成图像时遵循了对象之间的关系
3) 具有生成复杂图像的能力。
4) 模型预测的布局可能与真实的布局有差距，有时候这会成为模型的瓶颈

作者还进行了如下的模型简化测试，结果如图 4所示，使用Inception分数作为评估指标：

- 省略图的卷积：因此可以从初始对象嵌入向量预测框和掩码。它不能联合推断不同对象的存在，只能预测每个类别中的一个框和掩码。

- 省略联系：使用图形卷积层，但是忽略输入场景图中的主要关系图形卷积允许这个模型联合处理对象。其较差的性能说明了场景图关系的实用性。
- 省略判别器：依靠像素回归损失来指导生成网络，往往会产生过于平滑的图像。

作者还对比了使用真实box和mask, 分别在有无图的卷积图情况下模型的性能。结果表明训练和测试时都使用真实box和mask，给出了模型性能的一个上限。即使在使用真实布局时，省略图卷积也会降低性能，这表明场景图关系和图卷积的好处不仅仅是预测目标位置。

一个图像生成模型效果的真正准确的评估方法应该是人为判别。于是作者将该生成模型与StackGAN对比进行人为评估，认为该模型生成的图像更贴近文本标题的人数超出StackGAN的两倍，而且评估者认为该模型生成的图像中所能识别出的物体是StackGAN的近1.5倍。

4. 结论

这篇论文提出了一种从场景图生成图像的端到端方法。与从文本描述生成图像的主要方法相比，从结构化场景图而不是非结构化文本生成图像允许模型可以显式地推理对象和关系，并能够生成具有许多可识别对象的复杂图像。

参考文献

[1] Han Z, Tao X, and Li H. StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks[j], 2016.

[2] Johnson J, Gupta A, and Fei-Fei L. Image Generation from Scene Graphs[J], 2018.