

摘要

文本生成图像作为近几年的热门研究领域，其解决的问题是从一句描述性文本生成与之对应的图片。近一周来，我通过阅读了近几年发表于顶会的近 10 篇论文，做出本文中对该方向的简要报告。报告中主要阐述了近几年最流行的解决方案——以 GAN 思想为主干的解决方案。首先我对现有方法进行了简单回顾，之后针对这些方法做出了自己的总结，将各方法中用来提升生成效果的方式归纳为“增加网络深度”、更加充分地利用文本信息及通过增加额外约束三种。继而又提出当前方法存在的不足以及自己对今后如何改进的简单思考。

1. 简介

从文本生成图像是近几年的热门研究领域，其主要任务是从一句描述性文本生成一张与文本内容相对应的图片。主流方法有 VAE(Variational Auto-Encoder)，DRAW (Deep Recurrent Attention Writer) 以及 GAN 等，其中 GAN 在近几年的研究中成为了最热门的方法，在大部分顶会论文中都用到了 GAN 的思想来完成图像的生成工作。无论使用何种 GAN，都先对自然语言文本进行处理得到文本特征，进而以该文本特征来作为后续图片生成过程的约束。在 GAN 中生成器 Generator 根据文本特征生成图片，继而被鉴别器 Discriminator 鉴定其生成效果，根据鉴别器的鉴定结果生成器再次生成更真实的图片，鉴别器则再次对新图鉴定，以此类推，迭代进行直到网络收敛。

2. 现有方法回顾

在 2016 年以前，VAE 和 DRAW 方法都被用来完成图像生成工作，VAE 以一种统计方法进行建模最大化数据的最小可能性来生成图像，而 DRAW 方法使用了循环神经网络，并利用注意力机制，每一步关注一个生成对象，依次生成一个 patch 并叠加出最终结果。其中 Mansimov, Elman, et al [3] 提出的 AlignDRAW 在传统 DRAW 的基础上加入了文本对齐，从而完成了文本到图像的任务。如图 1，该模型使用一个双向循环神经网络(BiRNN)作为文本编码器（图 1 左），将文本信息从正反两个方向编码为一个文本向量特征(text embedding)用于后面 DRAW 部分的文本对齐，DRAW 部分又有两部分构成，Inference 和 Generative，Inference 部分从输入图片和文本特征中逐步生成隐藏信息给 Generator，Generator 又从隐藏信息和对齐文本特征中每次一个 patch 地逐步生成图片。

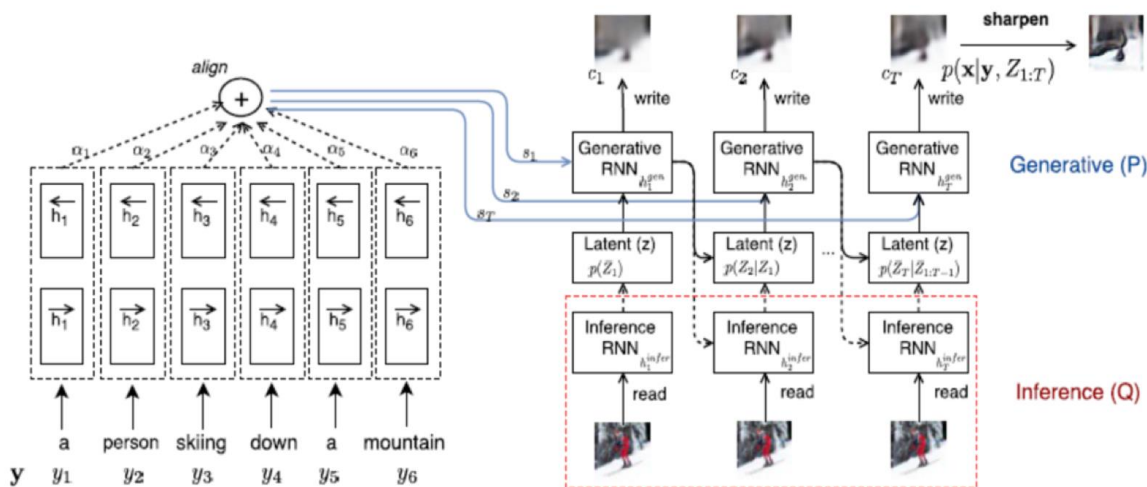


图 1. AlignDRAW 整体结构图

自 Reed et al [1] 2016 年提出 GAN-INT-CLS 以来，大部分的方法都使用了 GAN 的思想完成从文本到图像的任务。

GAN-INT-CLS 网络以 GAN 为模型主干（如图 2），同时在输入中增加文本特征来作为生成器和鉴别器的约束，最终生成 64x64 的图像。在生成器中，text embedding 跟随机噪声融合后一起输入到生成网络中；在鉴别器中，生成图像在下采样之后，跟之前的 text embedding 在空间复制之后融合，最后鉴别器根据融合特征进行判定。其中 GAN-CLS 主要加入了 Matching-aware discriminator，即在鉴别器中对错误情况进行分类(pair loss)，一种是生成的 fake 图像匹配了正确的文本，另一种是真实图像但匹配了错误文本，利用这种机制使得鉴别器网络不仅能够识别图像是否是生成器生成的(image loss)，并且能够鉴别生成图像跟给定文本的匹配关系，从而保证生成图像符合文本描述。GAN-INT 主要解决了文本信息的稀疏问题，在给出的文本特征中插值以获得生成图像的多样性。

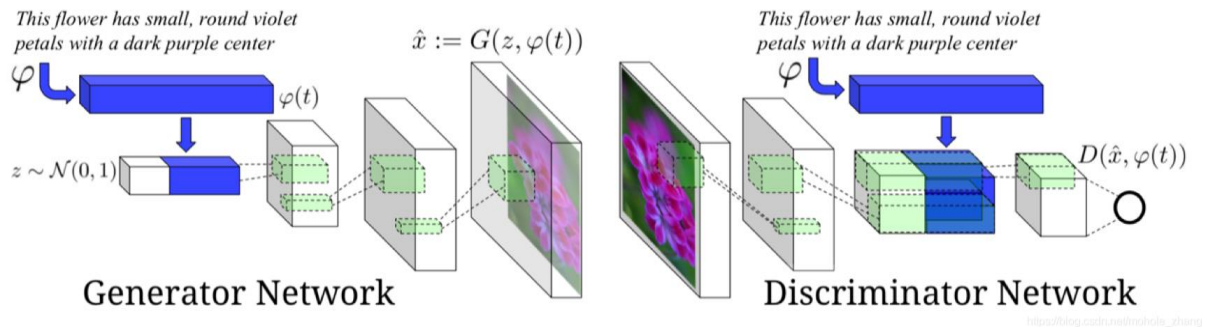


图 2. GAN-INT-CLS 总体网络结构

继 GAN-INT-CLS 之后，Reed et al 又在后续工作中给出了 Generative Adversarial What-Where Network(GAWWN)[2]，该文通过在网络中增加 bounding box 和 keypoint 限定，从而使生成图像精度提高，得到了 128x128 的图像。

StackGAN[4]在此基础上更进一步，使用了两个 GAN 来分步生成图像。因为单纯在网络中增加 up sampling 并不能提升生成图片的质量，所以 Zhang et al 提出了这样一个分两阶段的 GAN 网络，第一阶段用于生成低精度（64x64）的图像，该阶段主要关注图像的背景，颜色及轮廓等基本信息；在第二阶段中将第一阶段的输出作为输入同时再次使用 text embedding，从而获得了第一阶段丢失的细节信息，进而生成了 256x256 的更精细图片。同时在该方法中还加入了 CA(Conditioning Augmentation)模块来对文本特征加入一些实用的随机噪声，从而使得生成图像具有更多的可变性。在其后续工作中提出的 StackGAN++[5] 更进一步，将 GAN 扩充成一个树状的结构，采用了多个生成器和多个鉴别器并行训练，得到不同精度的图像（64x64, 128x128, 256x256），低精度生成器输出的隐层信息一方面用来生成低精度图，另一方面作为更高精度生成器的输入。在训练过程中各个生成器和鉴别器共享同一个 text embedding 保证了逐步提取更加精细的文本信息。同时该方法不仅可以完成限定性的生成任务(conditional generative tasks)，同时也扩展到了非限定性生成任务（unconditional generative tasks）。

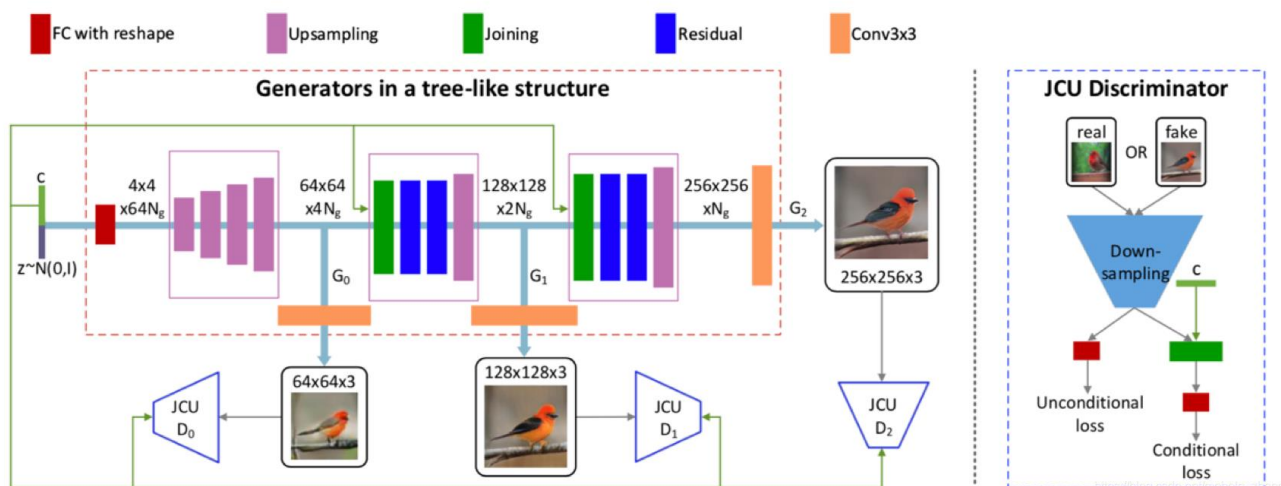


图 3. StackGAN++ 总体框架

之后，Xu et al 又在 StackGAN++ 基础上提出 AttnGAN [6]，相比 StackGAN++ 该方法增加了注意力机制，不仅提取文本的 sentence feature 作为全局约束，同时也将 attention 精确到 word 级别提取了 word embedding 作为局部约束送入网络，生成器与鉴别器每次针对 word embedding 部分精准优化，从而使得生成图像更能突出文本中的细节。此外，该文中还提出了一种 DAMSM (Deep Attentional Multimodal Similarity Model) 机制，该机制改进了训练过程中计算 loss 的方式，不仅考虑鉴别器的 source loss (即普通 GAN 的 loss)，同时更加关注训练过程中对 word embedding 的生成效果，在生成高精度图像后提取该图片的局部特征 (local image features) 跟 word embedding 进行对照进而获得 DAMSM loss 使得模型训练更加关注文本细节的生成情况，从而使得生成效果得到了提升。

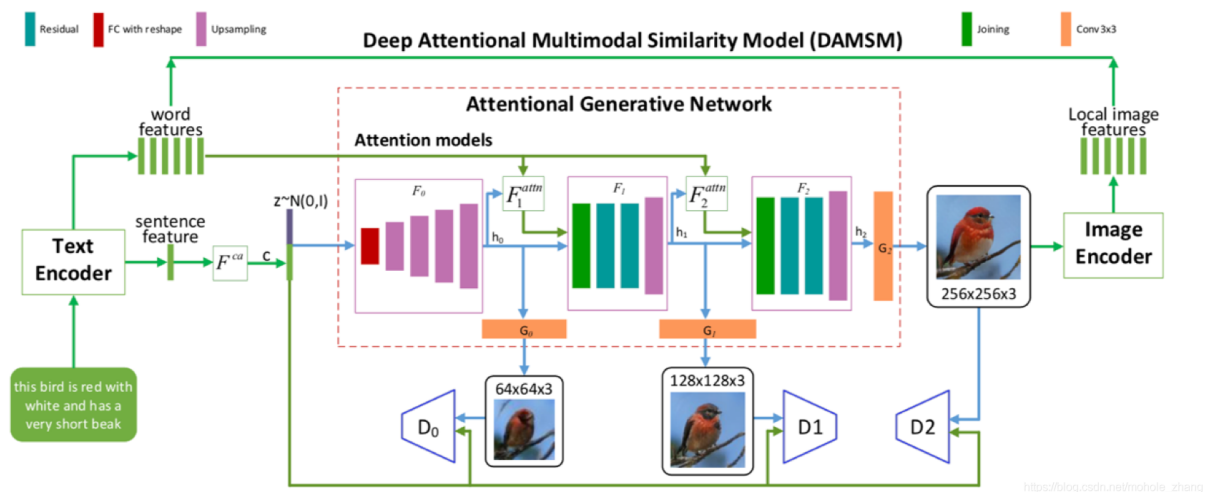


图 4. AttnGAN 总体框架

Dash et al [7] 提出的 TAC-GAN 借鉴了 AC-GAN 的思想，不仅考虑了文本约束，同时增加

了 class label 作为约束，在鉴别器中鉴别结果不止有原来的鉴定生成图片或真实图片，同时也鉴定其类别信息，通过增加类别约束提升了生成效果。

Johnson et al [8]则不在 text embedding 的维度进行约束，而是更深入到文本语义，提出了通过 scene graph 来建模文本中各对象及其关系，在获得 scene graph 的基础上对语义中的每个对象得到其 bounding box 和 mask 进而得到一个关于文本语义的 scene layout，然后以此 scene layout 作为输入加入到后续的 GAN 网络中生成图片。该方法改善了之前各种方法难以生成复杂场景的问题，使得生成图像更能反映文本语义。

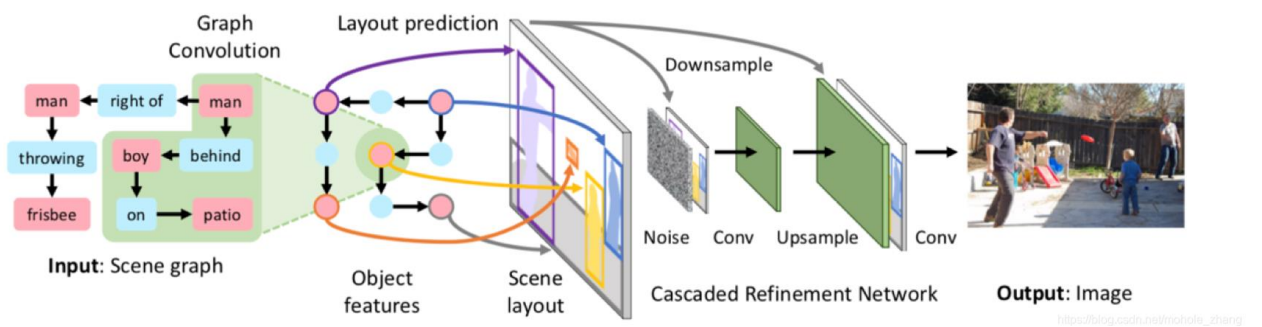


图 5. 利用 scene graph 生成图像的总体框架

Hong, Seunghoon, et al [9]提出了 HDGAN 借鉴 StackGAN 系列的思想，但是只使用了一个统一的生成器(single-stream Generator)同时带有多个级联鉴别器的网络模型，实现了端到端方式的图像生成，并且无需 class label 等额外的约束信息。Hong, Seunghoon, et al [10] 提出的"Inferring Semantic Layout for Hierarchical Text-to-Image Synthesis." 中将文本生成图像过程分为两步，首先从文本到语义框架(text to semantic layout)然后再生成图像，在从文本到语义框架过程中，又分为两步，先从文本中通过 LSTM 网络获得各个对象实例的 bounding box 然后利用 BiLSTM 在每个实例对象的 bounding box 中预测对象实例的语义 mask 然后将 bounding box 和 mask 结合成为 semantic layout 作为后续 GAN 的输入。

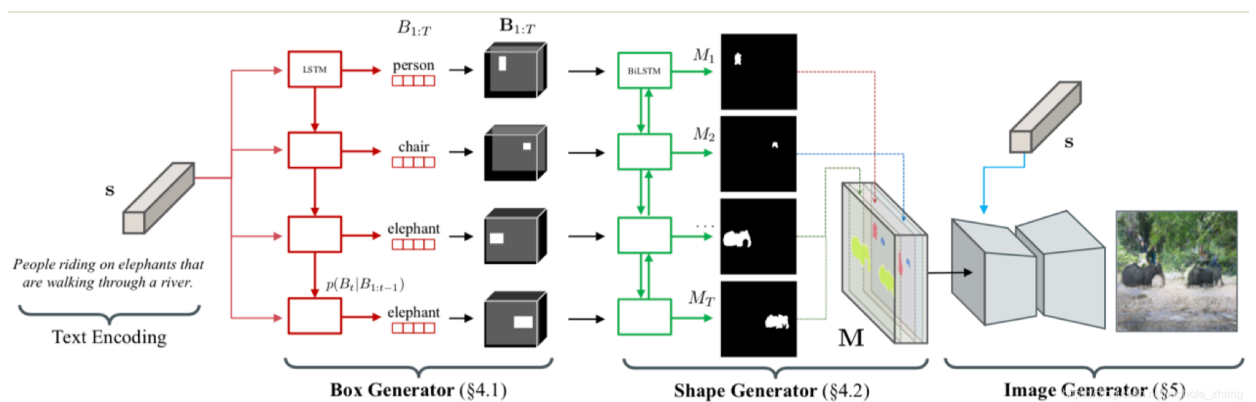


图 6. "Inferring Semantic Layout for Hierarchical Text-to-Image Synthesis."中 semantic layout 的生成过程

3. 现有方法总结

对比自 2016 年到 2018 年顶会中各主流文本到图像生成的方法可以看出，GAN 的思想几乎被所有方法用到，表明 GAN 在图像的生成中的确有着明显的优势。再深入到各个方法分析，我个人将各种方法对生成质量优化的方式分为三种：第一通过“增加网络深度”；第二通过更加充分地利用文本信息；第三通过增加额外约束。

对于“增加网络深度”，从最初的 GAN-INT-CLS[1]和 GAWWN[2]仅使用一个 GAN，到 StackGAN 系列 ([4,5,6]) 及 HDGAN[9]用到两个甚至多个 GAN 进行训练，如果我们把焦点放在网络整体过程上，可以说从一个 GAN 到多个 GAN 的过程正是将网络加深的过程，从低精度到高精度这样一个金字塔结构，文本特征和图像特征在网络中走过了更多的“层”，这种方式使得网络能够提取更多文本信息，避免单个 GAN 遗失信息的问题，从而可以得到更好的生成效果。

而针对更充分的利用文本信息，在早期的各个方法中 ([1],[2],[4]) 更多的只是将注意力放在对生成图像部分的网络进行调整，而对于文本的关注则只是关注其总体特征，利用已有的文本处理模型将文本处理成 text embedding(sentence embedding)，可以说其粒度还是相对较粗的，能够关注到的细节不够多。而在 AlignDRAW[3]及后续工作 ([5],[6]) 中则进一步将文本信息挖掘到更细粒度的 word 级别，通过更细粒度的约束来提升生成效果；最后，在最近使用的 scene graph[8]及 semantic layout[10]中则更进一步，不仅考虑细粒度的文本特征，同时还对文本语义进行更深入的挖掘，做到了对文本中实例及其关系的建模，从而将这些语义信息反映到一个 layout 中间层上，使得模型能够处理更加复杂的场景。

对于通过增加额外约束的方式，几乎所有方法中都有涉及。其中最为明显的是各个方法都使用了额外的 loss 来优化，例如在 GAN-INT-CLS[1]和 GAWWN[2]中提到的两种 loss (image loss, pair loss) 被大多数后来的方法所借鉴；在 StackGAN 系列中生成器的训练中加入 KL 正则项 ($DKL(N(\mu(t), \sigma(t)) \parallel N(0, I))$)；而在 TAC-GAN[7]中则是加入了 class loss；在 Image Generation from Scene Graphs [8]中则更是综合考虑了六种 loss (Box loss, Mask loss, Pixel loss, Image adversarial loss, Object adversarial loss, Auxiliary classifier loss)。

版权声明：本文为 CSDN 博主「mohole_zhang」的原创文章，遵循 CC 4.0 BY-SA 版权协议，转载请附上原文出处链接及本声明。

原文链接：https://blog.csdn.net/mohole_zhang/article/details/89374420