基于 GAN 的文本图像生成

慕宗燊
11821060
浙江大学计算机学院
cszsmu@163.com

摘要 (Abstract)

最近随着 GAN 在生成类任务取得不错的进展，一些新颖的工作例如文本生成图片（或视频）的任务（图像描述的逆任务）受到越来越多的关注，它对于艺术生成、计算机辅助设计、肖像生成、剧本生成电影等目标有重要应用。本文总结了近期文本生成图像的进展和最新工作，我们将工作分成单帧和多帧图片生成两类分析了其创新性和算法的主要流程，以及它们的主要实验结果。这些模型对于图片生成在细粒度和高分辨率方面做出了贡献，对于视频生成在一致性和连贯性方面得到了提升，与现有主流算法相比结果更好。

1. 背景

文本描述自动生成图像是许多应用中的基本问题，例如艺术生成和计算机辅助设计。这是近年来最活跃的研究领域之一，它推动了多模态学习和跨媒体（视觉和语言）推理的研究进展。文本生成图像是建立本文和图像的连接，打通语义鸿沟，通过此能够模拟人脑中文本到图像生成的过程以及解释人类产生星象的原因和过程。

随着深度学习在分类问题上取得了重大突破，人们希望能过将他同样运用到生成类任务当中，先前有许多基于图像或者视频生成相关文本描述的工作 [1][2][3][4]，他们大多数都是先将视觉信息通过卷积神经网络编码然后送入循环神经网络解码得到相关描述信息，在此基础上实现了很多变体，生成的描述通过评判标准进行训练，因为每个人对视觉信息有不同的理解，生成描述结果的语义多样性、连也只是差强人意。但是从视觉信息到语义信息的逆问题——文本生成图像（或视频）对于现有方法来说是自由度更高的挑战任务，对于文本到图像的映射不同的个体又不一样的先验知识对物体的客观印象造成脑中的星象不唯一也不尽相同，这类任务的评判标准也更难衡量和制定。

目前大多数文本到图像生成方法都是基于生成对抗网络（GAN）实现的，将整个文本编码得到一个句子向量并作为生成图像的条件，通过判别器和生成器之间的博弈训练得到生成图像。本文总结了 2018 年以来 AAAI 和 CVPR 中典型的文本生成图像（或视频）的最新进展，包括两篇生成单帧图片的 AttnGAN[5] 和 HDGAN[6]，两篇生成多帧图片的 GTT2V[7] 和 StoryGAN[8]，这些文章取得了令人印象深刻的结果进一步推动了生成模型的发展。AttnGAN 主要是解决条件 GAN 的输入是一个全局句子向量造成生成的图片缺乏细节的控制阻碍了高质量图片生成的问题，它结合注意力机制多阶段细粒度生成图片；HDGAN 主要是解决 GAN 生成图像像素质量较小，在生成器中间层多尺度嵌入判别器加入了深度分层的对抗约束，使从文本到图像的输出更平缓而不是跳跃的；GTT2V 创新性的提出利用文本生成短帧视频，利用文本先生成一个背景图再由背景图结合文本生成细腻度的动作变换；StoryGAN 则是解决生成短帧视频的连贯性问题，通过一个类似 RNN 网络每个单元生成一句话的图片，多句话生成具有连贯性和一致性的短视频。

2. 相关工作

变分自编码（VAE）、生成对抗网络（GAN）和基于流（flow-based）的模型已经广泛应用到各种生成任务当中，变分自编码从实验生成的效果来看存在模糊的问题，流模型具有较强的解释性但是因为生成和判断是两个可逆的过程因此在时间上存在效率问题，目前主流的方法还是以 GAN 为基础的各种变体结构运用到文本生产图像（或视频）的任务中，GAN 生成的图片更犀利有更强的随机性。尽管 GAN 取得了重大进展，但仍存在许多未解决的困难，例如训练不稳定和高分辨率生成，最近也有许多工作通过改变结构和引入正则化致力于解决上述问题。[9]

文本到图像生成的关键任务是理解输入文本和学习从低维流行到复杂的实际图像分布的连续映射。Reed 等人首先证明了条件 GAN 是能够从文本描述中生成图像的可能性 [10]，随后通过引入额外知识例如物体标签、位置等可以增强样本生成的效果 [11]。

与文本到视频生成最相关的工作是给定一张图像生成视频，Chen 等人提出了基于静态图片生成短视频的方法，其关键方法在于将潜在的移动物体与给定的图像区分开来实现动作的变化生成 [12]。对于文本到视频的工作没有现有的第一帧因此很直接的想法是利用 GAN 生成文本到视频的第一帧图片。

受这些工作的启发，我们讨论的几篇文本生成图像（或视频）的工作改进了不稳定性模型，生成了高分辨

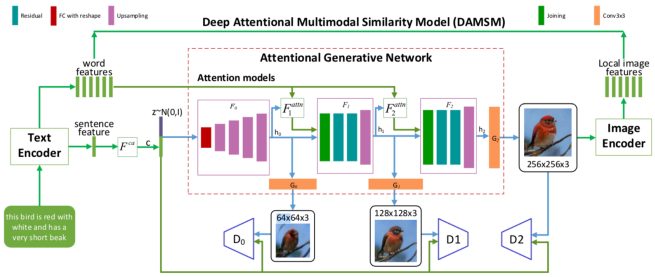

Figure 1. AttnGAN 的模型架构，注意力部分生成图片的不同子区域，DAMSM 提供细粒度的文本图像匹配损失

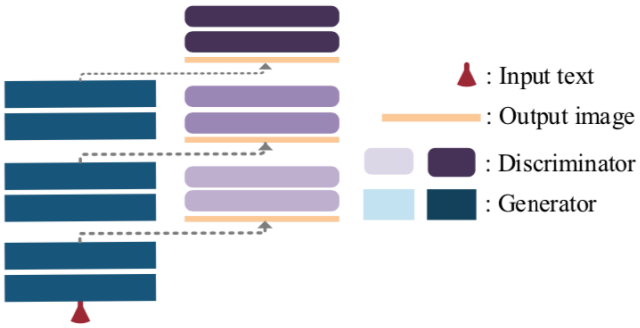

Figure 2. HDGAN 上采样生成高质量图片过程中结合分阶段判断器的端到端模型架构

率图像，实现了视频的连贯性和一致性。

3. 算法

本文将文本信息到视觉信息生成任务分为文本到单帧和多帧图片生成两类，多帧图片生成的任务和单帧任务类似，只是增加对图片连贯性和一致性的要求，单帧任务的方法可以迁移到多帧任务中。

3.1. 单帧图片的生成

AttnGAN 的整体框架如图 1 所示，包括了注意力模块和深度注意力多模态相似性模块（DAMSM）两部分。其中注意力网络包括 m 个生成器 $(G_0, G_1, ..., G_{m-1})$，其中隐藏层状态 $(h_0, h_1, ..., h_{m-1})$ 作为输入，生成有小到大规模的图片区域 $(\hat{x_0}, \hat{x_1}, ..., \hat{x_{m-1}})$。公式表示如下：

$$\begin{aligned} h_0 &= F_0(z, F^{ca}(\bar{e})); \\ h_i &= F_i(h_{i-1}, F_i^{attn}(e, h_{i-1})) \ for \ i = 1, 2, ..., m; \\ \hat{x}_i &= G_i(h_i). \end{aligned} \quad (1)$$

其中 z 表示从标准正态分布采样的随即向量，$\bar{e}$ 表示一个句子编码后的全局向量，e 表示句中单词组成的向量矩阵，F^{ca} 表示将句子向量 $\bar{e}$ 映射到条件向量的条件转换函数，F_i^{attn} 表示每个阶段的注意力子模块，通过神经网络训练 $F^{ca}, F_i^{attn}, F_i, G_i$。注意力模块计算权重的方法和运用在其他任务上的类似隐藏层状态 h 作为查询 e 作为值乘积后进行 softmax 得到每个单词在当前子区域生成中占的比重。深度注意力多模态相似性模块学习将图像的子区域和句子的单词映射到一个共同的语义空间两个神经网络，从而测量单词级别的图像-文本相似度，以计算图像生成的细粒度损失。整个模型训练的损失函数由所有生成子区域的模块损失 $\mathcal{L}_G = \sum_{i=0}^{m-1} \mathcal{L}_{G_i}$ 和相似度度量模块的损失 $\mathcal{L}_{DAMSM}$：

$$\mathcal{L} = \mathcal{L}_G + \lambda \mathcal{L}_{DAMSM}, \ where \ \mathcal{L}_G = \sum_{i=0}^{m-1} \mathcal{L}_{G_i}, \quad (2)$$

λ 用来平衡两个模块的表现，通过后向传播更新参数学习生成网络和判别网络的参数。

HDGAN 使用了和 AttnGAN 类似的思路，在逐步生成最终图片的过程分阶段的判断低维度的中间图片是否和原始感受野中的图对应。较低分辨率的输出用于学习语义一致的图像结构（例如对象草图、颜色和背景），随后的高分辨率侧面输出用于渲染细粒度的细节。由于我以端到端的方式进行训练，较低分辨率的输出也可以充分利用来自较高分辨率鉴别器自上而下的知识。因此可以保证在低分辨率和高分辨率图像的输出中观察到一致的图像结构、颜色和样式。与 StackGAN 相比，它通过堆叠两个低分率到高分辨率的生成器分别训练，难以保证两部分输出的一致性。

HDGAN 探索了沿着生成器的深度玩对抗性游戏的新维度（如图 2 所示），在生成器的中间层集成了额外的分阶段嵌套判别器，其充了生成器隐藏空间的正规化器，这也为误差信号流提供了一条短路径，并有助于减少训练的不稳定性。生成器 $\mathcal{G}$ 是一个上采样的 CNN 其每一步产生的输出为：

$$X_1, ..., X_s = \mathcal{G}(t, z) \quad (3)$$

其中 $t \ p_data$ 表示一个句子的向量，$X_1, ..., X_{s-1}$ 是分辨率逐渐递增的中间图片，X_s 是最后得到的最高分辨率图片。对于每一次生成器的输出 X_i 都有一个与之对应的判别器 D_i，最后生产对抗的规则就是 min-max 游戏寻找最优生成器和判别器：

$$\mathcal{G}^*, \mathcal{D}^* = \arg\min_{\mathcal{G}} \max_{\mathcal{D}} \mathcal{V}(\mathcal{G}, \mathcal{D}, \mathcal{Y}, t, z), \quad (4)$$

其中 $\mathcal{Y} = Y_1, ..., Y_s$ 表示在 s 个生成阶段中分辨率相对应的实际图片。为了保证图像的语义信息一致性和图像的保真度，每个判别器需要判别两部分损失。全局语义损失由文本和图像对的一致性衡量，局部细腻度的保真损失由生成器每个阶段不同感受野中像素级别的真假判断。

3.2. 多帧图片的生成

GTT2V 整个模型框架（图 3）包含三个重要成分：条件的背景主题生成器、视频生成器和视频判别器。背景生成器是利用条件变分自编码框架，将视频中第一帧图片作为输入然后通过编码器编码到隐藏空间，结合文本向量进行采样通过解码器得到背景，因为 VAE 模

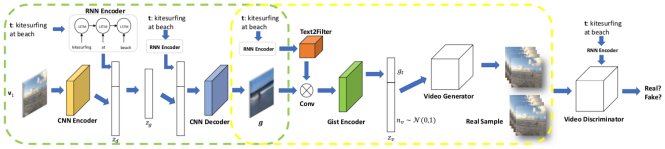

Figure 3. GTT2V 包括 Gist 生成器（绿色部分）、视频生成器（黄色部分）和视频判别器

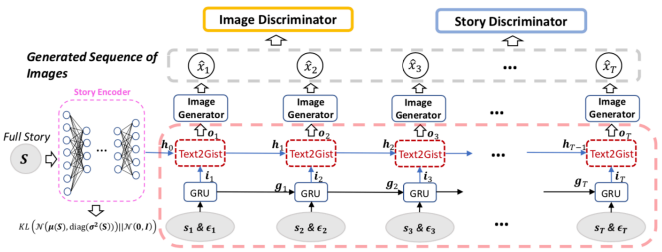

Figure 4. StoryGAN 的整体框架，最下面是故事和句子的编码器，中间是图片生成器，最上面是两部分判别器

型生成的样本与原始图片通过散度距离衡量所以为了损失更小，生成的样本大多较为模糊，我们利用此作为整个生成视频的基调。视频生成器模块主要的贡献就是结合了文本滤波器（Text2Filter）来对背景图提取语义信息，通过高效的卷积操作将文本和视觉空间的信息融合，经过上采样得到生成的视频。利用视频判别器判断视频与原始视频的差距。整个过程通过无监督的学习训练三部分神经网络，该模型创新性的提出文本到视频的生成过程，利用首张图片定基调然后文本信息进行图片细节的生成。

StoryGAN 给定长度为 T 句话 $S = [s_1, s_2, ..., s_T]$ 的故事，产生与之对应的图片序列 $\hat{X} = [\hat{x}_1, \hat{x}_2, ..., \hat{x}_T]$。真实的图片序列为 $X = [x_1, x_2, ..., x_T]$，我们希望生成的图片和真实图片能够保持局部和全局的一致性，局部一致体现为每张图片能和每句话语义信息相对应，全局一致性体现在所有图片和整个故事保持协调。整个故事向量表示为 S，每句话的向量为 s_i，整个架构如图 4 所示，包括了将故事向量 S 编码到低维空间 h_0 的编码器；一个两层循环神经网络作为故事内容编码器在时间 t 将每句话 s_t 和故事梗概内容 h_t 编码得到主旨 o_t；一个图像生成器基于 o_t 生成 $\hat{x}_t$；最后是一个图像判别器和故事判别器分别保证了局部和全局的一致性。

其中内容编码器包含两部分：

$$i_t, g_t = GRU(s_t, \epsilon_t, g_{t-1}) \tag{5}$$

$$o_t, h_t = Text2Gist(i_t, h_{t-1}) \tag{6}$$

GRU 提炼每句话的局部内容信息，h_{t-1} 为前一个时间单元得到的全局信息，将两部分结合捕获当前时间点的图片主旨 o_t，将它作为图片生成器的输入。图像判别器测量所生成的图像 x_t 是否与给定在初始上下文信息为 h_0 下的句子 s_t 匹配，即比较

	In-set	DT2V	PT2V	GT2V	T2V
Accuracy	0.781	0.101	0.134	0.192	0.426

Table 1. GTT2V 模型在 Youtube 测试集上真实结果和消融实验的准确率

	StotyGAN vs ImageGAN	
Choice(%)	StoryGAN	ImageGAN
Visual Quality	74.17±1.38	18.60±1.38
Consistence	79.15±1.27	15.28±1.27
Relevance	78.08±1.34	17.65±1.34

Table 2. StoryGAN 和 ImageGAN 在三个指标上人类评估结果

三元组 $\{s_t, h_0, \hat{x}_t\}$ 和 $\{s_t, h_0, x_t\}$ 的真假；故事判别器是将生成的图像序列通过编码得到一个特征向量 $E_{img}(X) = [E_{img}(x_1), ..., E_{img}(x_T)]$，整个故事编码得到一个特征向量 $E_{txt}(X) = [E_{txt}(s_1), ..., E_{txt}(s_T)]$ 衡量这两个向量的相似度得到全局的一致性分数：

$$D_S = \sigma(w^T(E_{img}(X) \odot E_{txt}(S) + b) \tag{7}$$

由该分数得到基于整个故事生成的视频是否为真。

4. 实验

AttnGAN[1]、HDGAN[2]、GTT2V[3]和 StoryGAN[4]均提供了源代码，可以参考更多实现细节。

图 5 分别展示了 AttnGAN 和 HDGAN 生成模型的可视化结果，AttnGAN 结果的第一排展示了不同生成器 G_i 生成的图片结果，第二排和第三排分别是 G_0 和 G_1 在文本注意力下前五个重要单词关注的子区域热点图，可以看出模型较准确的捕捉了单词和图像区域的相似性；HDGAN 的结果从上到下分别显示了三个样本在生成高清图片中不同分辨率图片。

因为文本生成视频的可视化结果不便于展示可以参考 github 中相关可视化的结果，表 1 展示了 GTT2V 模型和自身消融实验在测试集上的准确度，DT2V 直接将文本作为输入利用 GAN 生成视频，PT2V 在判别器损失上设计了判断文本和视频对的误差，GT2V 不使用文本过滤器，T2V 结合所有模块，四组消融实验可以看出和人工标注的 0.781 准确度相比，模型的各个子模块均起作用达到了 0.426 的准确度。StoryGAN 和 ImageGAN 的实验对比结果如表 2 所示，在视觉质量、单张图片的一致性和多帧图片的相关性上利用人类评估可以看到 StoryGAN 取得了突破式的提升，说明模型在生成视频中保持了局部和全局一致性。

[1]AttnGAN: https://github.com/taoxugit/AttnGAN
[2]HDGAN: https://github.com/ypxie/HDGan
[3]GTT2V: https://github.com/yitong91/text2video
[4]StoryGAN: https://github.com/yitong91/StoryGAN

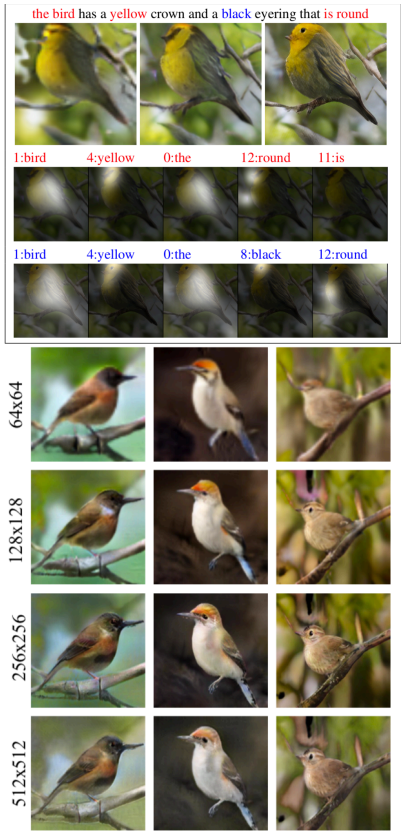

Figure 5. 上半部分是 AttnGAN 结果，下半部分是 HDGAN 结果

5. 总结

本文概述了四种有文本信息生成视觉信息的生成模型，它们展示了 GAN 及其变体在生成类模型中的强大应用。但是从实验结果我们可以看到，生成的图像或视频并不是很理想，整体看着有物体的轮廓其中细节还是很突兀缺少约束。最近英伟达也推出基于迁移风格的 GAN 生成图像补全，谷歌也推出了模型复杂度计算度都庞大的 GAN 生成模型，这些模型虽然在高清图或者图像大致轮廓取得了进展，但是由于缺少人类的先验知识或者知识库的引导，还是很难捕捉细节。这或许也是深度学习的不可解释性造成模型生成图像的不可控性，因此要做到更满意的生成模型未来还有很多值得探讨的研究。

参考文献

[1] Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T. and Saenko, K. 2015. Sequence to sequence-video to text. In IEEE ICCV.

[2] Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K. and Darrell, T. 2015. Long-term recurrent convolutional networks for visual recognition and de- scription. In IEEE CVPR.

[3] Pan, Y., Mei, T., Yao, T., Li, H. and Rui, Y. 2016. Jointly modeling embedding and translation to bridge video and language. In IEEE CVPR.

[4] Pu, Y., Min, M. R., Gan, Z. and Carin, L. 2017. Adaptive feature abstraction for translating video to language. ICLR workshop.

[5] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In CVPR, 2018.

[6] Zhang, Z., Xie, Y., Yang, L. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In CVPR, 2018.

[7] Y. Li, M. R. Min, D. Shen, D. Carlson, and L. Carin. Video generation from text. In AAAI, 2018.

[8] Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson and Jianfeng Gao.StoryGAN: A Sequential Conditional GAN for Story Visualization. In CVPR, 2019.

[9] T.Salimans,I.Goodfellow,W.Zaremba,V.Cheung,A.Radford, and X. Chen. Improved techniques for training gans. In NIPS, 2016.

[10] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text-to-image synthesis. In ICML, 2016.

[11] A. Dash, J. C. B. Gamboa, S. Ahmed, M. Z. Afzal, and M. Liwicki. Tac-gan-text conditioned auxiliary classifier generative adversarial network. arXiv preprint arXiv:1703.06412, 2017.

[12] Chen, B.; Wang, W.; Wang, J.; Chen, X.; and Li, W. 2017. Video imagination from a single image with transformation generation. In ACMMM workshop, 2017.