评述 Semantics Disentangling for Text-to-Image Generation

白耘宇 21821244
浙江大学
浙大路 38 号
yunyubai@zju.edu.cn

Abstract

CVPR 2019 中关于对抗生成网络的 *oral paper* 有两篇，而 *Semantics Disentangling for Text-to-Image Generation* 是其中唯一的 *text-to-image* 的论文；同时该文的方法更是大幅提升了图片生成的效果 *(*在 *MS COCO* 数据集上，提升了 *Inception Score* 大约 *10* 个百分点！*)*。这是本人推荐的两个理由。接下来会介绍这篇论文的特色、创新性求解思路、重要的实验结论。

1. Introduction

现有的 text-to-image 的工作 [1, 6] 主要聚焦在提升图片的质量和分辨率，或者是利用堆叠起来的“由粗糙到精致”的生成器结构，或者是通过采用注意力机制的生成过程。然而，这些方法都忽视了一个问题，那就是人类对于一张图片的描述是非常主观且因人而异的，所以简单地利用这些文本作为唯一的描述，来生成图片常常会导致生成不稳定，即具有相似语义但是表述方式不同的描述会生成差别较大的图片。

为了解决这个问题，作者提出了 SD-GAN，可以从描述中提取出一致性的语义信息，同时保留描述的多样性和其中的细节信息。具体地，作者使用孪生网络结构来保留图片和描述之间的语义一致性；并通过条件批量归一化的方法来保留文本的多样性。

1.1. Siamese structure network

Siamese network 最早在 2005 年 Yann Lecun 提出，它可以看做是一种相似性度量方法，衡量两个输入的相似程度。它的模型架构如图 1 所示，其中左右两个网络共享参数，因此可以看作时完全相同的网络。

如果只是在像素空间中进行相似性度量显然不合适，因此 Siamese network 通过某个映射函数将输入的样本映射到一个目标空间中，然后在目标空间中使用一般的距离度量方式进行相似度比较，希望同类的样本的距离应该相近，而不同类别的样本距离应较远。$G_W(X)$ 表示需要学习的映射函数，在孪生网络中只要求其可微，并不附加其它的限制条件，参数 W 的求解是主要工作。对于输入 X_1 和 W_2，当它们是同类别的样本时，相似性度量 $E_W(X_1, X_2) = \|G_W(X_1) - G_W(X_2)\|$ 的值较小；当它们是不同类别的样本时，相似性度量 $E_W(X_1, X_2) = \|G_W(X_1) - G_W(X_2)\|$ 的值较大。因此，在训练集上使用成对的样本进行训练，输入同类别时最小化损失函数 $E_W(X_1, X_2)$，而输入不同类别时最大化损失函数。

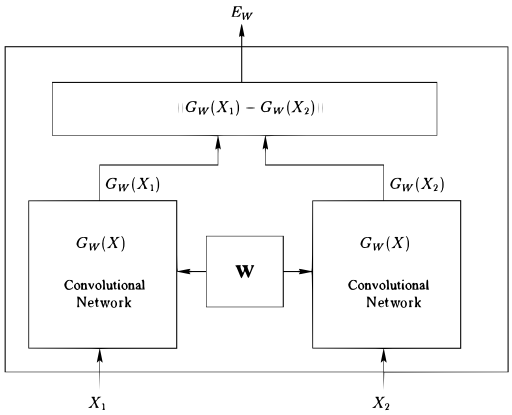

Figure 1. Siamese structure network

1.2. Conditioned Batch Normalization

神经网络学习过程本质就是为了学习数据分布，一旦训练数据与测试数据的分布不同，那么网络的泛化能力也大大降低。

输入层的数据，已经人为的的归一化，后面网络每一层的输入数据分布是一直在发生变化的，前面层训练参数的更新将导致后面层输入数据分布的变化，因此必然会引起后面每一层输入数据分布的改变。而且，网络的前面几层发生微小的改变，那么后面几层就会被累积放大下去。我们把网络中间层在训练过程中，数据分布的改变称之为：”Internal Covariate Shift”。批归一化的提出，就是要解决在训练过程中，中间层数据分布发生改变的情况。

而条件批量归一化是对批量归一化的改进，可以将其看作是在一般的特征图上的缩放和移位操作的一种特例，

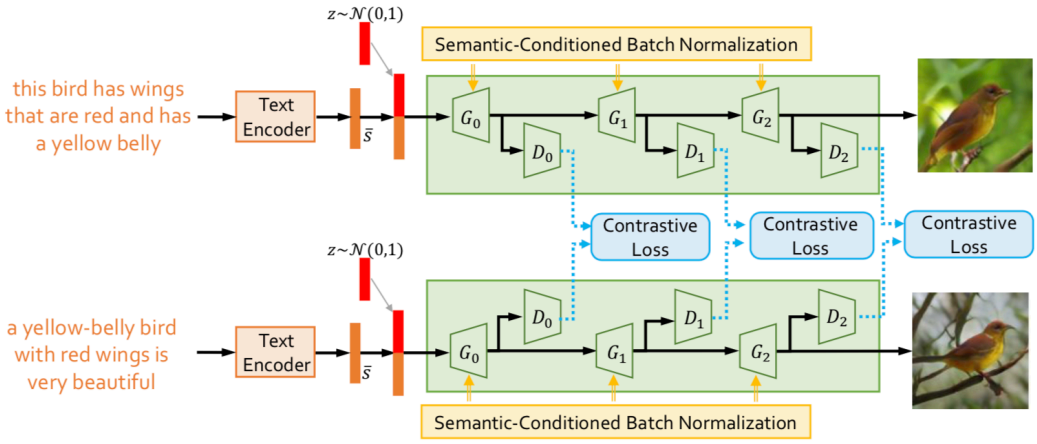

Figure 2. SD-gan 架构.

2. Semantics Disentangling Generative Adversarial Network(SD-gan)

本文提出了一个新颖的跨模态生成式网络——语义分离式对抗生成网络 (SD-gan)，来解决 text-to-image 的生成，模型架构如图 2 所示。主要创新点如下：(1) 利用孪生网络结构，使得生成的图片不仅仅依赖于当前网络分支上的输入信息，同时也受到另一个网络分支的输入信息影响。换句话说，孪生网络结构提取了文本描述的共同语义来解决文本表述多样化的问题；(2) 为了提高图片的生成质量，SD-gan 又需要保留输入文本不同的细节信息。为此作者提出了语义条件的批归一化 (SCBN)。

2.1. Siamese Structure with Contrastive Losses

前文已提到，SD-gan 采用孪生网络提取文本语义信息来实现跨模态生成。作者采用了对比损失，最小化同一张图片下两种文本描述所生成的假图片之间的距离，同时最大化不同的真图片对应的文本所生成的假图片的距离。在训练阶段，生成的图片受到来自两个网络分支的文本的控制。

如图 2 所示，孪生网络的每个分支网络都采用了堆叠起来的生成器-判别器模块，包括：1) 从文本描述中提取文本特征的文本编码器；2) 多阶段的生成对抗子网，既有图像生成器 G_0, G_1, G_2，又有对应的判别器 D_1, D_2, D_3。

文本编码器。每个网络分支的输入都是一句自然语言描述。文本编码器旨在从自然语言描述中学得特征表示，与 [8, 9] 一样，作者采用了双向的 LSTM 来提取语义向量。一般来讲，在双向 LSTM 中，每个隐藏状态都表示了句子中一个单词的语义，而最后一个隐藏状态则被作为全局的句子向量 $\bar{s}$。

Hierarchical Generative Adversarial Networks. 参考了 [8, 9]，本文也使用从低分辨率到高分辨率的多个阶段来生成图片。基于文本编码器得到的 $\bar{s}$ 和从标准正态分布采样得到的噪音向量 z，在初始阶段得到分辨率 64×64 的图片，如图 3(a) 所示（其中的 SCBN 层将在下节介绍）。下一个阶段则使用前一阶段的输出和句子向量 $\bar{s}$ 来产生更高分辨率的图片，如图 3(b) 所示。在每一个阶段，生成器之后都跟着一个判别器来判断图片是否真实。判别器 D_1, D_2, D_3 是独立的，并不共享参数。

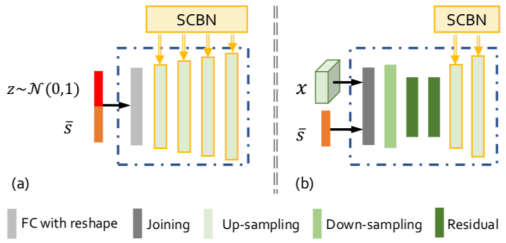

Figure 3. SD-gan 中的生成器:(a)G_0, 初始阶段的生成器，从语言到图片；(b)G_1/G_2，第二/第三阶段的生成器，用于生成更高分辨率的图片 SCBN 层在每一个上采样层之后进行操作

Contrastive Loss. 作者采用孪生网络的目的是提高生成图片在语义上的一致性。对孪生网络的两个分支输入两种不同的文本描述，如果两个分支生成的图片是语义相似的，那么生成图片也应该相似；反之，生成图片也应该不同。本文作者最后采用了对比损失来提取文本对的语义信息。

对比损失起初是由 [4] 提出的，且损失函数是

$$L_c = \frac{1}{2N}\sum_{n=1}^{N} y \cdot d^2 + (1-y)max(\epsilon - d, 0)^2 \qquad (1)$$

这里 $d = \|v_1 - v_2\|_2$ 是两个分支的视觉特征向量 v_1, V_2 之间的距离，y 是标签，表示输入的描述是否来自同一张图片，比如，1 表示相同，0 表示不同。超参数 N 是特征向量的长度，在本文的实验中凭经验取 256。

超参数 ϵ 用于平衡当 $y=0$ 时的距离，在实验中设置为 1。

通过最小化来自同一图片的两段描述所生成的假图片之间的距离，以及最大化来自不同图片的两段描述所生成的假图片之间的距离，孪生网络结构得以优化。当然，由于输入噪声的存在，哪怕是两段描述是完全一样的，最后生成的图片也或多或少有所不同。而为了避免模式崩溃的出现，作者修改了 (1) 式，使得特征向量之间的最小距离不为 0。修改后的式子如下：

$$L_c = \frac{1}{2N}\sum_{n=1}^{N} y \cdot max(d,\alpha) + (1-y)max(\epsilon - d, 0)^2 \quad (2)$$

实验中的 α 设置为 0.1，这样即使两段描述是来自同一张图片，最后生成的假图也不会距离太近。

2.2. Semantic-Conditioned Batch Normalization (SCBN)

作者认为，对于跨模态的 text-to-image 生成，语言上的概念是视觉表示的关键。参考过去归一化的工作 [2, 5]，作者利用自然语言描述中的语言学线索，修改了条件批归一化，并将其命名为语义条件批归一化 (SCBN)。SCBN 的作用是强制把视觉语义嵌入到生成网络的特征图中。这样就使得语义嵌入可以控制视觉的特征图，比如可以扩大或缩小特征图，对其取反等等。作者认为这样就弥补了孪生网络仅仅提取共同语义的不足。

批归一化给定一批输入 $x \in R^{N\times C\times H\times W}$，BN 归一化每一个通道的均值和方差如下：

$$BN(x) = \gamma \cdot \frac{x-\mu(x)}{\sigma(x)} + \beta \quad (3)$$

其中 $\gamma, \beta \in R^C$ 是从数据学得的参数，$\mu(x), \sigma(x)$ 是沿着 batch 的维度计算得到的均值和标准差，并且每个特征通道是独立的。

条件批归一化与 BN 只学得单独的一套参数不同，[3] 提出了条件批归一化，可以学得在条件线索 c 下的修正参数 γ_c, β_c。CBN 是更广泛的特征图放缩变换操作的特例。修改后的归一化函数如下：

$$BN(x|c) = (\gamma + \gamma_c) \cdot \frac{x-\mu(x)}{\sigma(x)} + (\beta + \beta_c) \quad (4)$$

语义条件批归一化为了强迫视觉语义的嵌入，作者在生成器中加入了提出的 SCBN 层，如图 3 所示。首先，作者从词编码器中得到输入描述的特征。不妨记第 t 个单词的特征为 w_t，最后的隐藏状态为全局句子向量 $\bar{s}$。因此，SCBN 的语言学线索来自句子级别和词级别两个方面。

(1) 句子级别线索。为了嵌入句子特征，作者采用单隐层的多层感知器 (MLP) 从句子向量 $\bar{s}$ 中提取修正参数 γ_c,β_c，如图 4(a) 所示。公式如下：

$$\gamma_c = f_\gamma(\bar{s}), \beta_c = f_\beta(\bar{s}) \quad (5)$$

这里的 $f_\gamma(\cdot)$ 和 $f_\beta(\cdot)$ 分别对应 γ_c,β_c 的 MLP。然后，作者分别延展 $f_\gamma(\bar{s})$ 和 $f_\beta(\bar{s})$ 的维度大小，使其与 x 相同，从而使用式 (4) 将语言学线索嵌入到视觉特征中。

(2) 单词级别线索。不妨记 $\mathcal{W} = \{w_t\}_{t=1}^{T} \in R^{D\times T}$ 是一句描述对应的单词特征，w_t 是第 t 个单词的特征，$\mathcal{X} \in R^{C\times L}$ 是对应的视觉特征，其中 C 是通道的大小，而 $L = W \times H$。参考 [7]，作者使用了视觉语义嵌入 (VSE) 模块来融合单词特征和视觉特征，如图 4(b) 所示。作者先是使用一层感知器 ($f(w_t)$) 使得词向量的维度与视觉特征维度匹配。接着就开始按照下式 (6) 计算 VSE 向量 vse_j。

$$vse_j = \sum_{t=0}^{T-1} \sigma(v_j^T \cdot f(w_t)) f(w_t) \quad (6)$$

其中，v_j 是图像的第 j 个子区域的特征，而 $\sigma(v_j^T \cdot f(w_t))$ 指的是第 t 个词向量 w_t 对于视觉特征图子区域 v_j 的权重，累加起来即得到的 vse 向量。在作者的实验中，$\sigma(\cdot)$ 采用 softmax 函数。最后，基于 VSE 矩阵，使用两个 $conv_{1\times1}$ 来分别计算单词级别的修正参数 γ_c,β_c。

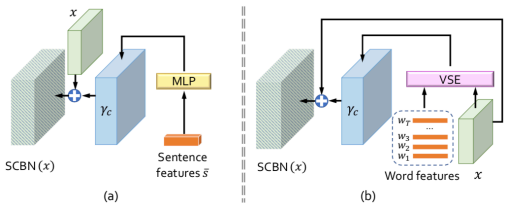

Figure 4. SCBN.(a) 基于句子向量，通过单隐层 MLP 提取修正参数的句子级别线索；(b) 利用 VSE 模块融合视觉特征和词特征的单词级别线索。图示是举 γ_c 为例，β_c 的实现与之相似

3. Experiments

3.1. Experiments Setting

数据集. 作者将 SD-gan 在 CUB 数据集和 MS COCO 数据集上进行了评估。CUB 数据集共有 11788 张图片，包含了 200 种鸟类，每张图片对应 10 份文本描述。作者将 CUB 数据集划分为类别分离的训练集和测试集，即 8855 张训练、2933 张测试；并对 CUB 所有的图片进行了预处理和剪切，保证鸟类的检测框占图片面积的 75% 以上。而 MS COCO 更有挑战性，它的训练集有 8 万张图，测试集有 4 万张，每张图片对应 5 份文本描述。

评估细节. 作者采用了两种评估方式。一种是定量的方式，使用评估 GAN 常用的 Inception Score；同时由于 Inception Score 仅仅反映图片质量，不能说明生成图片是否与文本内容一致，作者采用了另一种定性的方法对模型进行评估：给定文本描述，运行 SD-gan 和其它 gan 得到生成图片，而有 50 个用户为不同模型的生成图片排序，最后计算每个模型有多少张图片被认为是最佳图片，并算出其所占比例。

3.2. Comparing with the state-of-the-arts

作者将 SD-gan 与其它先进的模型在 CUD 和 MS COCO 数据集上进行了比较。SD-gan 和其它 gan 的 Inception Score 如图 5 所示。可以发现，不仅仅在 CUB 数据集上，SD-gan 取得了最优的结果；更是在 MS COCO 数据集上，提升了 Inception Score 大约 10 个百分点！这令人惊叹的结果，证明了作者提出的语义分离生成的优越性。

而定性比较的结果如图 6 所示，SD-gan 与 StackGAN 和 AttnGAN 进行了比较，其中 SD-GAN 生成的图片被认为是最优的比例均在 70% 左右。

Methods	CUB	MS-COCO
GAN-INT-CLS [29]	$2.88 \pm .04$	$7.88 \pm .07$
GAWWN [30]	$3.62 \pm .07$	-
StackGAN [40]	$3.70 \pm .04$	$8.45 \pm .03$
StackGAN++ [41]	$4.04 \pm .05$	-
PPGN [24]	-	$9.58 \pm .21$
AttnGAN [37]	$4.36 \pm .03$	$25.89 \pm .47$
HDGAN [42]	$4.15 \pm .05$	$11.86 \pm .18$
Cascaded C4Synth [19]	$3.92 \pm .04$	-
Recurrent C4Synth [19]	$4.07 \pm .13$	-
LayoutSynthesis [14]	-	$11.46 \pm .09$
SceneGraph [18]	-	$6.70 \pm .01$
SD-GAN	$\mathbf{4.67 \pm .09}$	$\mathbf{35.69 \pm .50}$

Figure 5. SD-gan 与其它方法的定量比较

Methods	CUB	MS-COCO
StackGAN [40]	10.70%	6.53%
AttnGAN [37]	20.54%	17.69%
SD-GAN	68.76%	75.78%

Figure 6. SD-gan 与其它方法的定性比较

4. Conclusion

SD-GAN 有效利用了输入文本的语义来生成图片。它采用了孪生网络结构从文本描述中提取共同语义，从而可以在表述方式多变的情况下保持生成图片的一致性。而仅仅使用孪生网络结构可能会丢失某些唯一的语义，故 SD-GAN 又采用了一种改善的视觉语义嵌入方法来保留输入文本不同的细节信息。最终在 MS COCO 和 CUB 上的实验也证明了，SD-GAN 确实具有非凡的性能。

References

[1] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. *CoRR*, abs/1612.05424, 2016.

[2] Harm de Vries, Florian Strub, Jérémie Mary, Hugo Larochelle, Olivier Pietquin, and Aaron C. Courville. Modulating early visual processing by language. *CoRR*, abs/1707.00683, 2017.

[3] Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. *CoRR*, abs/1610.07629, 2016.

[4] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742, June 2006.

[5] Xun Huang and Serge J. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. *CoRR*, abs/1703.06868, 2017.

[6] Jianan Li, Xiaodan Liang, Yunchao Wei, Tingfa Xu, Jiashi Feng, and Shuicheng Yan. Perceptual generative adversarial networks for small object detection. *CoRR*, abs/1706.05274, 2017.

[7] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. *CoRR*, abs/1711.10485, 2017.

[8] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaolei Huang, Xiaogang Wang, and Dimitris N. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *CoRR*, abs/1612.03242, 2016.

[9] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. Stackgan++: Realistic image synthesis with stacked generative adversarial networks. *CoRR*, abs/1710.10916, 2017.