武汉大学学报(信息科学版)
Geomatics and Information Science of Wuhan University
ISSN 1671-8860,CN 42-1676/TN

《武汉大学学报(信息科学版)》网络首发论文

题目：行列式点过程采样的文本生成图像方法
作者：李晓霖，李刚，张恩琪，顾广华
DOI：10.13203/j.whugis20210373
收稿日期：2021-06-21
网络首发日期：2022-07-18
引用格式：李晓霖，李刚，张恩琪，顾广华．行列式点过程采样的文本生成图像方法[J/OL]．武汉大学学报(信息科学版). https://doi.org/10.13203/j.whugis20210373

武 汉 大 学 学 报 · 信 息 科 学 版
Geomatics and Information Science of Wuhan University

DOI:10.13203/j.whugis20210373

引用格式：李晓霖，李刚，张恩琪，等. 行列式点过程采样的文本生成图像方法[J]. 武汉大学学报 · 信息科学版，2022，DOI：10.13203/j.whugis20210373（Li Xiaolin, Li Gang, Zhang Enqi, et al. Determinant Point Process Sampling Method for Text-to-Image Generation[J]. *Geomatics and Information Science of Wuhan University*, 2022, DOI: 10.13203/j.whugis20210373）

行列式点过程采样的文本生成图像方法

李晓霖 [1,2]　李刚 [1,2]　张恩琪 [1,2]　顾广华 [1,2]

1　燕山大学 信息科学与工程学院, 河北　秦皇岛, 066004

2　河北省信息传输与信号处理重点实验室, 河北　秦皇岛, 066004

摘　要：近年来，基于生成对抗网络(Generative Adversarial Networks, GAN)的文本生成图像问题取得了很大的突破，它可以根据文本的语义信息生成相应的图像。然而目前生成的图像结果通常缺乏具体的纹理细节，而且经常出现模式崩塌、缺乏多样性等问题。本文针对以上问题，提出一种针对生成对抗网络的行列式点过程方法(Determinant Point Process for Generative Adversarial Networks, GAN-DPP)来提高模型生成样本的质量，并使用 StackGAN++、ControlGAN 两种基线模型对 GAN-DPP 进行实现。在训练过程中，该方法使用行列式点过程核矩阵对真实数据和合成数据的多样性进行建模，并通过引入无监督惩罚损失来鼓励生成器生成与真实数据相似的多样性数据，从而提高生成样本的清晰度及多样性，减轻模型崩塌等问题，并且无需增加额外的训练过程。在 CUB 和 Oxford-102 数据集上，通过 Inception Score、Fréchet Inception Distance 分数、Human Rank 三种指标的定量评估，证明了 GAN-DPP 对生成图像多样性与质量提升的有效性。同时通过定性的可视化比较，证明使用 GAN-DPP 的模型生成的图像纹理细节更加丰富，多样性显著提高。

关键词：生成对抗网络；文本生成图像；行列式点过程；模型崩塌；多样性

近年来，深度学习技术迅速发展，并在模式识别与分类[1]、图像修复[2]、图像生成[3]等领域得到广泛应用，其中，文本图像生成任务是一个重要研究方向。文本生成图像允许使用自然语言来描述视觉概念，在图像生成领域提供了灵活的人机交互功能，具有极大的应用潜力，例如图像编辑[4]、漫画生成[5]等。文本生成图像问题，以自然语言文本作为输入，生成包含文本信息的图像，并保证图像的视觉质量。这是一种跨模态[6][7]问题，与计算机视觉和自然语言处理[8]两个方向密切相关，计算机在对文本语义进行理解的同时也需要生成与描述匹配的图像。

实际生活中，针对同样的文本描述，人对图像的设想都会有所不同，但都与文本描述的语义相匹配，这就涉及到文本生成图像的一对多特性。比如给出句子“一个女孩正在看书”，对于图像中女孩的发型、衣着等细节，情况可以是多种多样的，因此符合文本的图像结果不止有一种模式，这就是多样性问题。针对有限文本发掘更多符合语义的图像模式，提升生成样本的多样性，有利于充分发挥文本生成图像模型的作用，提高文本生成图像任务的应用效果。

2014 年，Goodfellow 等人提出了生成对抗网络(Generative Adversarial Networks, GAN)[9]，广泛应用在计算机视觉的各个领域，尤其是图像生成问题。条件生成对抗网络（Condition Generative Networks，cGANs）[10]在 GAN 的基础上加入外部信息作为附加输入，如：类标签[11]、文本[12]、图像[13]等，可以约束生成过程，使 GAN 相关问题的解决有了很大突破[14]。Reed 等人[12]于 2016 年提出了 GAN-CLS 模型，首次利用 GAN 网络根据文本在 Oxford-102[15]和 CUB[16]数据集上生成了 64*64 的图像，但生成图像像素较低且缺乏细节，易出现过拟合现象。Zhang 等人[17]提出了堆叠生成对抗网络（StackGAN）生成 256*256 的高分辨率图像，但其两阶段的模型训练方式导致了模型训练计算难度的增加；随后 Zhang 等人[18]提出了 StackGAN++，通过树状结构来解决多阶段训练的问题。但方法仍存在一系列问题，例如生成图像的细节

收 稿 日 期：2021-06-21

项 目 资 助：国家自然科学基金(62072394)、河北省自然科学基金(F2021203019)。

第 一 作 者：李晓霖，硕士生，研究方向为生成对抗网络、文本生成图像。imlixlin@163.com

通 讯 作 者：顾广华，博士，教授，主要研究方向为跨模态信息处理。guguanghua@ysu.edu.cn

不够精细、多样性不足等。MSGAN[19]针对生成对抗网络训练中的模式崩塌[20]现象，提出采用正则化项来进行解决，改善图像结果多样性不足的现象，但是效果有限，生成样本质量与多样性仍旧有待提升。

Xu 等人提出的 AttnGAN[21]在生成对抗网络中引入注意力机制，通过单词与图像子区域的相关性推动模型生成图像不同子区域上的细粒度特征，通过多阶段训练生成细粒度图像。Li 等人提出了 ControlGAN[22]，在 AttnGAN 的空间注意力基础上引入通道注意力获得单词和视觉特征中各通道的关联性，减少来自不相关通道的影响，以此生成具有细粒度特征的高质量图像。一些算法虽然生成的图像质量有了提高，但是受数据集约束，生成样本仍然缺乏多样性。

现有的大部分文本生成图像方法，仅注重图像真实性而忽略文本生成图像所具有的一对多特性，导致生成图像集中于少数几种模式而缺乏多样性。针对此问题，本文从数据样本的多样性建模入手，通过行列式点过程[23]采样实现建模过程。本文提出生成对抗网络的行列式点过程方法(GAN-DPP)，采用分辨率逐级提升的模型实现图像生成，并在训练中采用提出的行列式点过程损失。采用 DPP 核矩阵实现对生成与真实数据的建模，通过提出的无监督损失推动生成器生成样本的多样性接近真实数据的多样性。

本文通过定性与定量实验，对 Oxford-102 与 CUB 数据集上的生成结果进行评估与可视化比较。结果表明，GAN-DPP 通过引入行列式点过程核矩阵与无监督惩罚损失，使文本生成图像模型在保证高质量图像生成的同时，在图像多样性生成方面的性能也得到有效提升。GAN-DPP 模型更好地实现了从文本到图像的生成，在跨模态图像生成领域具有较高的应用价值。

1 基于 GAN-DPP 的生成对抗网络

文本生成图像的研究大部分借助于生成对抗网络，其中堆叠式生成对抗网络能够实现由低分辨率至高分辨率的图像生成。而引入注意力机制能够将文本单词与图像子区域对应，提高图像细节部分生成质量。

1.1 堆叠生成对抗网络

堆叠对抗网络结构具有不止一对的生成器与判别器，在网络的树状结构中，每个分支都有一个生成器生成相应分辨率的图像，并且在同一分支中有对应的判别器对图像特征进行判别。

图 1 为堆叠生成对抗网络结构，模型采用预训练编码器（char-CNN-RNN）[24]对给出的文本 t 进行编码，得到对应的特征向量 φ_t。由于文本特征向量维度较高，进行简单的非线性转换可能会产生不连续数据流，因此本文采用条件增强[18]方法来获得低维文本向量 t'。将来自标准正态分布的噪声 $z\sim P_{noise}$ 和条件变量 t' 拼接作为输入，并且经过逐层网络生成隐藏特征 h_i：

$$h_0=F_0(z,t'),\quad h_i=F_i(h_{i-1},t') \tag{1}$$

其中 h_i 为网络结构第 i 个分支的输出隐藏特征，设网络的分支数量为 a，有 i=1,2，…，a-1。将不同分支的隐藏向量 h_i 作为相应阶段生成器的输入，于是能够得到第 i 个分支的生成图像结果 $S_i=G_i(h_i)$。

生成器 G_i 生成图像后，将真实图像 x_i 和生成图像输入到判别器中。每个分支的判别器由两部分组成，一部分判断输入的图像是否真实，另一部分判断图像与文本描述是否匹配，D_i 和 G_i 交替训练，直到收敛。

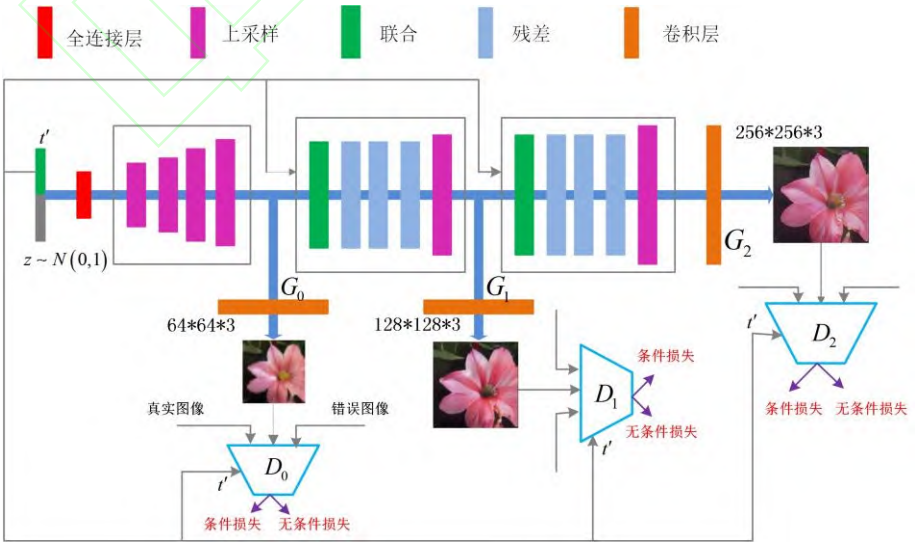

图 1　堆叠生成对抗网络结构图

Fig.1　Stacked Generative Adversarial Network Structure Diagram

堆叠生成对抗网络通过生成器 G_0 生成 64*64 分辨率的图像，生成器 G_1、G_2 则生成 128*128 与 256*256 分辨率的图像。判别器主要由下采样层与卷积层构成，将相应阶段的图像与文本条件作为输入，并转化为 4*4*512 的向量，并且根据损失函数进行计算。

1.2 注意力生成对抗网络

文本生成图像模型将整个文本句子编码为一个向量，会导致文本细粒度信息缺失。由此注意力生成对抗网络引入注意力模块，通过计算单词与图像子区域的相关性，将与子区域最相关的单词作为约束条件，以此提升图像细节质量，网络结构如图 2 所示。

为获得文本编码，注意力生成对抗网络文本编码器采用双向长短期记忆网络(Bi-LSTM)[25]从文本描述中获得句子与单词向量。模型将噪声 $z\sim P_{noise}$ 与经过条件增强方法得到句子向量 t' 作为输入，e 为经过文本编码器得到的单词向量，F_i^{attn} 为注意力模块。

对于网络的 a 个 $(G_0, G_1, \cdots, G_{a-1})$ 生成器，除 G_0 外，各分支的生成器以注意力模块生成的隐藏特征向量 $(h_1, h_2, \cdots, h_{a-1})$ 作为输入，生成分辨率从小到大的图像 $(s_0, s_1, \cdots, s_{a-1})$。注意力模块将单词特征转换到图像特征的公共语义空间得到向量 e'，并计算图像各子区域的词上下文向量 c_j，即各单词与图像特征的相关性表示：

$$c_j = \sum_{i=0}^{T-1} \beta_{j,i} e_i', \quad \beta_{j,i} = \frac{\exp\left(s'_{j,i}\right)}{\sum_{k=0}^{T-1} \exp\left(s'_{j,k}\right)} \tag{2}$$

有 $s'_{j,i} = h_j^T e_i'$，模块输出为 $F^{attn}(e,h) = (c_0, c_1, \ldots, c_{N-1})$，表示图像各个子区域与单词的相关性。注意力模块通过计算文本内单词与图像各子区域的相关性，使模型在生成图像细节时在相关语义上分配更多注意力，增强细粒度信息在图像生成中的作用，提高了图像细粒度特征的生成质量。

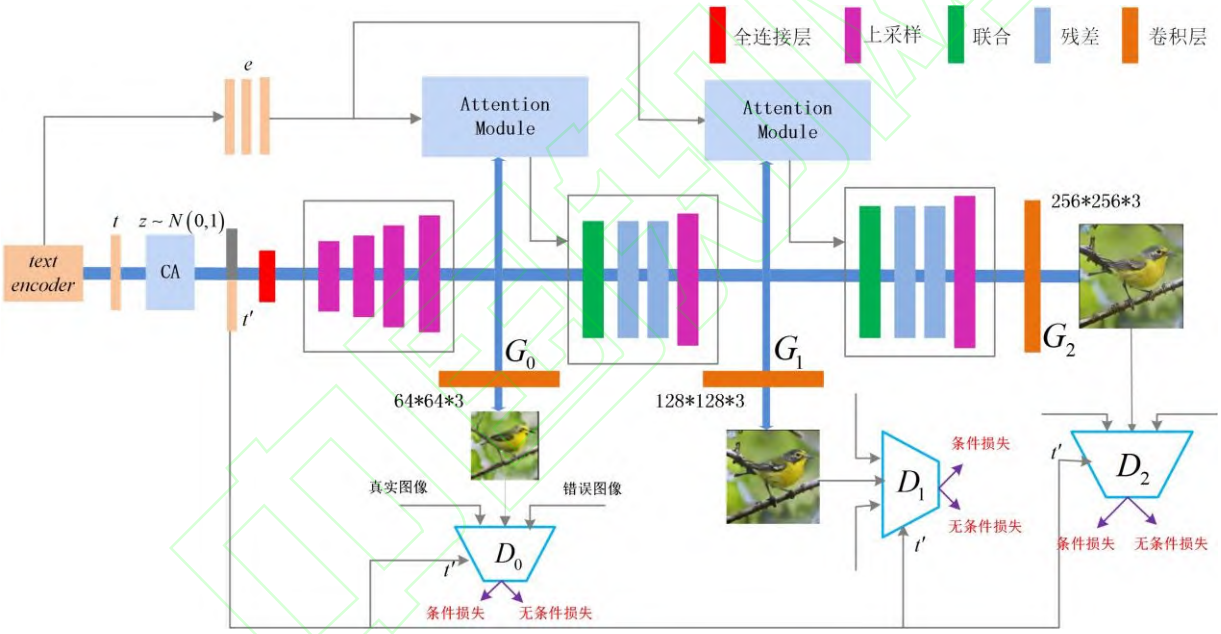

图 2 注意力生成对抗网络结构图

Fig.2 Attention Generative Adversarial Network Structure Diagram

如图 2 所示，单词向量与图像特征经过注意力模块所得到的表示 $F_i^{attn}(e, h_{i-1})$，与上一阶段的隐藏特征 h_{i-1} 结合得到隐藏特征向量 h_i，作为下一阶段的输入。网络通过生成器 G_0，G_1 和 G_2 分别得到分辨率为 64*64、128*128 和 256*256 的生成图像，并送入各分支判别器进行判别。

1.3 行列式点过程 DPP

行列式点过程 DPP[26]在量子物理领域被应用为概率度量来模拟高斯-泊松和费米尔过程[27]，随后于随机矩阵理论中得到较多研究。通过 DPP 能够较好地获取负相关性[28][29]来实现对相似性的衡量，因此能够实现对子集数据多样性的度量。由于 DPP 不了解一个分布子集中数据的顺序，可用其来建模给定分布中随机抽取的数据，例如对真实数据的训练集进行小批量采样。

设有集合 y，行列式点过程 P 能够在集合的幂集 2^y 上进行概率测量，使用 P 进行采样，其得到的子集可以是空集或全集。设通过 P 随机采样得到的子集为 Y，则对任意 $M \subseteq Y$，有：

$$P(M \subseteq Y) \propto \det(L_M) \tag{3}$$

DPP 的核矩阵是 $N*N$ 的实对称方阵，将其表示为 L。设子集 M 的相似性内核矩阵为 L_M，并且将其行列式值表示为 $\det(L_M)$。核矩阵 L 中，L_{ij} 表示的是集合

y 中样本之间的相似性，即 L_{ij} 越大，两个样本{i，j}的相似度越高，二者被同时采样的概率越低。

Kulesza 等人[30]提出将核矩阵 L_M 分解为葛朗姆矩阵(Gram-matrix)：

$$P(M \subseteq Y) \propto \det\left(\phi(M)^T \phi(M)\right) \prod_{e_i \in M} q^2(e_i) \quad (4)$$

式中，$q(e_i) \geq 0$ 为集合 y 中 e_i 项的质量分数，$\phi_i \in \mathbb{R}^D, \|\phi_i\|_2 = 1$ 是 e_i 项的 L2 正则化特征向量。有 $\det\left(\phi(M)^T \phi(M)\right) = \Pi_i \lambda_i$，$\lambda_i$ 为核矩阵 $\phi(S)^T \phi(S)$ 的第 i 个特征值。

想要采样所得样本更加丰富多样，可理解为采样点之间的距离相对更远，此时则不希望多个样本点聚集情况出现，此问题最简单的采样方式为均匀采样，其结果如图 3 左图所示，能够观察到仍旧存在多个点聚集，未达到足够均匀和全面的采样效果。图 3 右图对应为 DPP 采样情况，可以看出采样点之间距离较远，样本分布整体更加稀疏，进而多样性也得到了提升。

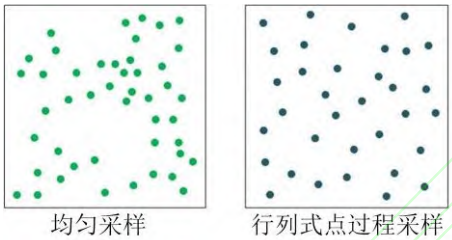

图 3　DPP 采样和均匀采样效果对比

Fig.3　Comparison of DPP Sampling and Uniform Sampling

1.4　点行列式过程方法

生成对抗网络判别器的损失如公式 5 所示，前两项无条件损失判断图像真假；后两项条件损失判断图像与文本语义是否匹配：

$$\begin{aligned} L_{D_i} = & -\mathbf{E}_{x_i \sim P_{data_i}}\left[\log D_i(x_i)\right] - \mathbf{E}_{s_i \sim P_{G_i}}\left[\log\left(1 - D_i(S_i)\right)\right] \\ & -\mathbf{E}_{x_i \sim P_{data_i}}\left[\log\left(D_i(x_i, t')\right)\right] - \mathbf{E}_{s_i \sim P_{G_i}}\left[\log\left(1 - D_i(S_i, t')\right)\right] \end{aligned} \quad (5)$$

对于生成器，损失函数同样包括有条件与无条件两项。前者用来训练网络生成更加真实的图像，后者训练网络生成与给出的文本条件相匹配的图像：

$$L_{G_i} = -\mathbf{E}_{s_i \sim P_{G_i}}\left[\log D_i(S_i)\right] - \mathbf{E}_{s_i \sim P_{G_i}}\left[\log\left(D_i(S_i, t')\right)\right] \quad (6)$$

本文提出的方法将网络中判别器 D_2 生成结果之前的隐藏层提取出，作为特征提取函数 $\phi(\bullet)$。在训练中加入行列式点过程损失，在减少冗余网络结构与计算量的同时，鼓励生成器生成图像的多样性提升。因此判别器 D_2 共有 3 个任务，分别为鉴别图像是否为生成图像、判断图像是否满足文本条件约束、以及计算行列式点过程损失，其结构如图 4 所示。

设生成模型 G 产生采样 $S_B = \{e_1, e_2, \cdots\cdots, e_B\}$，对输入噪声 z_B 和文本 t'，有 $S_B = G(z_B, t')$。对于每次

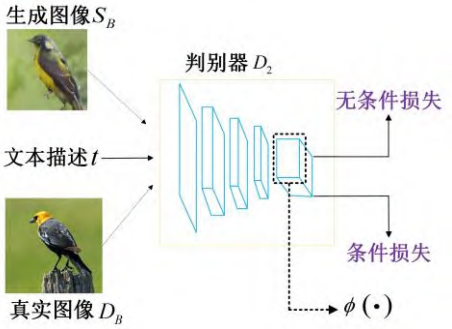

图 4　判别器结构图

Fig.4　Discriminator Structure Diagram

迭代，假设真实分布为 P_d，且有一批采样为 $D_B \sim P_d$。则对于 D_B 采样真子集，其 DPP 核矩阵表示为 L_{D_B}。主要目标为核矩阵 L_{D_B} 生成概率采样的 S_B 能够满足：

$$P(S_B \subseteq Y) \propto \det\left(L_{D_B}\right) \quad (7)$$

公式中 Y 是由点过程处理 P 生成的假子集的随机变量，通过公式 4 能够构建得到 S_B 与 D_B 分别对应的核矩阵 L_{S_B} 和 L_{D_B}。为合理控制变量，这里设真假子集采样的质量分数均为 $q(e_i) = 1$，于是可将公式简化为：

$$L_{S_B} = \phi(S_B)^T \phi(S_B), \quad L_{D_B} = \phi(D_B)^T \phi(D_B) \quad (8)$$

公式中 $\phi(S_B)$、$\phi(D_B)$ 分别为生成特征与真实特征，均由特征提取函数 $\phi(\bullet)$ 获得。GAN-DPP 方法学习假子集的 DPP 核矩阵 L_{S_B}，使其能够靠近真子集的 L_{D_B}，对于矩阵的匹配，可用具有的矩阵特征值与特征向量实现，即通过缩小匹配问题，实现矩阵特征值大小与特征向量方向的回归。

因此在 GAN-DPP 的损失函数包括多样性幅度损失 L_m 与多样性结构损失 L_s 两项，损失函数为：

$$\begin{aligned} L_{DPP} &= L_m + L_s \\ &= \sum_i \left\| \lambda_{real}^i - \lambda_{fake}^i \right\|_2 - \sum_i \hat{\lambda}_{real} \cos\left(v_{real}^i, v_{fake}^i\right) \end{aligned} \quad (9)$$

其中有 $\hat{\lambda}_{real}^i = \dfrac{v_{real}^i - \min\left(v_{real}^i\right)}{\max\left(v_{real}^i\right) - \min\left(v_{real}^i\right)}$，且 λ_{real}^i 和 λ_{fake}^i 分别对应为矩阵 L_{S_B} 和 L_{D_B} 的特征值，本文采用 $\hat{\lambda}_{real}^i$ 的最大最小归一化版本来约束特征向量 v_{real}^i 和 v_{fake}^i 的余弦相似性，以此减轻在实际分布中或训练过程中的噪声结构干扰。网络生成器与判别器的损失构成为：

$$L_G = \sum_i L_{G_i} + L_{DPP} + L_{KL} \quad (10)$$

$$L_D = \sum_i L_{D_i} \quad (11)$$

图 5 所示为生成器损失的组成结构，其中包括生

成损失 L_{G_i} 与多样性损失 L_{DPP}。生成损失鼓励生成器提升合成样本真实性，多样性损失鼓励生成器产生样本的多样性靠近真实样本多样性。

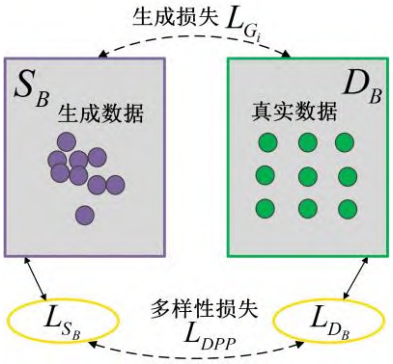

图 5　生成器损失组成图

Fig.5　Composition Diagram of Generator Loss

2 实验结果

2.1　实验环境和数据集

本实验过程中实验环境详细信息如下：

硬件环境：CPU：Intel(R) Core(M) i7-9700 @3.60GHZ；显卡：NVIDIA GeForce RTX2080Ti。

软件配置：操作系统为64位的Ubuntu18.04，CUDA Toolkit10.2，Python3.7，深度学习框架为Pytorch1.7。

本实验中，生成器和判别器的学习率都设置为0.0002，epoch 设置为 600，在 Oxford-102 数据集上批次大小为 16，CUB-200-2011 数据集的批次为 8。

本文采用的 CUB 数据集中包括 200 种鸟类，其中训练集与测试集分别具有 8855 与 2933 张图片，每张图片有 10 个文本描述。Oxford-102 数据集中包括 102 种类别的花，训练集与测试集分别具有 7035 与 1155 张图片，每张图片有 10 个文本描述。

2.2　评价指标

对于图像生成效果的评价，实验采用三种指标：Inception Score(IS)[31], Fréchet Inception Distance (FID)[32]，Human Rank(HR)[18]。

1) Inception Score

在图像生成领域，Inception Score 是一项常用指标，能够定量评估图像的质量与多样性。指标的公式如下：

$$IS = \exp\left(\mathbf{E}_x D_{KL}\left(P(y \mid x) \| p(y)\right)\right) \quad (12)$$

式中 x 为生成样本，y 为通过 Inception 模型[33]获得的预测标签。当模型生成的图像质量更优且多样性足够高时，则会有边缘分布 $p(y)$ 和条件分布 $p(x \mid y)$ 的 KL 散度相对更大。因此 IS 值更高，则代表图像的质量与多样性方面表现更优。

2) Fréchet Inception Distance

FID 的原理为通过 Inception 网络获取图像特征，进而在样本分布方向计算生成样本和真实样本之间的距离，通过参考真实数据对生成图像质量进行衡量评估，且效果与人类感知评估一致。指标计算公式如下：

$$FID = \left\|\mu_r - \mu_g\right\| + T_r\left(\Sigma_r + \Sigma_g - 2\left(\Sigma_r\Sigma_g\right)^{1/2}\right) \quad (13)$$

式中，μ_r 和 μ_g 为真实与生成图像特征的均值，Σ_r 和 Σ_g 为真实与生成图像特征的协方差。因此，FID 值越低时，则代表生成数据与真实数据之间越接近。

3) Human Rank

Human Rank 方法随机抽取图像与相应文本描述，通过参与者根据要求对图像进行排名，计算所得平均排名进行质量评估。指标包括两方向：整体质量等级(global quality rank, GQR)与局部质量等级(local quality rank, LQR)。GQR 为图像整体清晰度与背景丰富度评价，LQR 为图像与文本匹配程度评价。评价参与人员为随机选择的 20 人（不包括作者），对于打乱的四种模型生成图像，每人依据所指定方面对图像进行排名，效果由好到差的排名依次为 1，2，3，4。

2.3　实验结果

通过将本文 GAN-DPP 应用于堆叠生成对抗网络 StackGAN++与注意力生成对抗网络 ControlGAN，将模型分别命名为 GAN-DPP-S 与 GAN-DPP-C，并对引入 GAN-DPP 的模型生成结果进行定性和定量的评估。实验将采用了 GAN-DPP 方法的模型与文本生成图像领域已有模型分别进行了比较，包括 GAN-INT-CLS[12]、StackGAN[17] 、StackGAN++[18] 、AttnGAN[21] 、HDGAN[34]、MSGAN[19]以及 ControlGAN[22]。

文章采用 IS 与 FID 进行模型文本生成图像的定量评估，通过在生成图像数据中随机抽取 30000 个样本来实现相关分数的计算，表 1 和表 2 分别展示了 IS 与 FID 分数对比结果。通过使用 HR 分数与各模型生成结果的可视化进行模型定性评估，表 3 为 HR 分数对比，图 6、图 7 对比清晰度，图 8、图 9 对比多样性。

表 1 为不同模型在 CUB、Oxford-102 数据集上的 IS 分数对比，分数越高代表生成图像质量越高。在 Oxford-102 数据集上，GAN-DPP-S 的 IS 分数相较于 StackGAN++由 3.26 提高到 3.36，提高了 3.1%。对于 CUB 数据集，GAN-DPP-S 模型与 StackGAN++相比，IS 分数由 4.04 提高到 4.37，提高了 8.2%。针对注意力模型，GAN-DPP-S 的 IS 分数表现优于 ControlGAN。对于 Oxford-102 数据集，GAN-DPP-C 的分数由 3.81 提高到 3.94，提高了 3.4%。对于 CUB 数据集，GAN-DPP-C 分数由 4.53 提高到 4.62，提高了 1.9%。

表 1 不同模型在 Oxford-102 和 CUB 数据集上的 Inception Score 分数

Tab.1 Inception Score of Different Models on Oxford-102 and CUB Datasets

评价标准	数据集	GAN-INT-CLS	StackGAN	StackGAN++	GAN-DPP-S	AttnGAN	ControlGAN	GAN-DPP-C
IS↑	Oxford-102	2.66±.03	3.20±.01	3.26±.01	**3.36±.02**	—	3.81±.07	**3.94±.09**
	CUB	2.88±.04	3.70±.04	4.04±.05	**4.37±.04**	4.36±.03	4.53±.03	**4.62±.05**

表 2 不同模型在 Oxford-102 和 CUB 数据集上的 FID 分数

Tab.2 FID Scores of Different Models on Oxford-102 and CUB Datasets

评价标准	数据集	GAN-INT-CLS	StackGAN	StackGAN++	MSGAN	GAN-DPP-S	ControlGAN	GAN-DPP-C
FID↓	Oxford-102	79.55	55.28	48.68	—	**43.29**	28.22	**25.04**
	CUB	68.79	35.11	25.99	25.53	**24.33**	18.49	**17.90**

表 2 针对 CUB、Oxford-102 数据集展示了不同模型的 FID 分数，FID 数值越小，则说明模型生成的样本更加符合真实数据集表现，即图像质量越高。对于堆叠生成对抗网络，与 StackGAN++相比，GAN-DPP-S 在 Oxford-102 数据集上的 FID 分数由 48.68 下降到 43.29，降低了 11.1%；在 CUB 数据集上，FID 分数由 25.99 降低到 24.33，降低了 6.4%。由此可观察到，GAN-DPP-S 模型在 CUB 与 Oxford-102 数据集上的 FID 分数均有降低，即图像生成质量得到提高。对于引入注意力机制的网络，GAN-DPP-C 与 ControlGAN 相比，FID 分数得到了降低。其中对于 Oxford-102 数据集，GAN-DPP-C 的 FID 分数由 28.22 下降到了 25.04，降低了 11.2%，在 CUB 数据集上 GAN-DPP-C 的 FID 分数由 18.49 下降到了 17.90，降低了 3.1%。

表 3 不同模型在 Oxford-102 和 CUB 数据集上的 Human rank 分数

Tab.3 Human Rank of Different Models on Oxford-102 and CUB Datasets

数据集	评价指标	GAN-INT-CLS	StackGAN	StackGAN++	GAN-DPP-S
Oxford-102	GQR↓	3.80	2.50	2.05	**1.65**
	LQR↓	3.80	2.50	1.95	**1.75**
CUB	GQR↓	3.70	3.10	2.10	**1.10**
	LQR↓	3.95	3.00	2.05	**1.05**

表 3 比较了不同模型的 HR 分数，分数值低则代表排名更加靠前，即图像质量与语义一致性表现更优。对于 Oxford-102 数据集，GAN-DPP-S 的 GOR 分数为 1.65，LQR 分数为 1.75，在两种分数中均为最低，证明了在对比的所有模型中，GAN-DPP-S 能够达到更高的图像质量与语义匹配度。对于 CUB 数据集，GAN-DPP-S 的 GOR 分数为 1.10，LQR 分数为 1.05，低于其它模型的所得分数，说明采用 GAN-DPP 的模型在 CUB 数据集上依旧具有很好的表现，生成样本在图像清晰度与语义一致性方面均获得了提升。

图 6 为各模型在 Oxford-102 数据集上进行生成的可视化对比，可以看出 GAN-INT-CLS 只能够生成符合

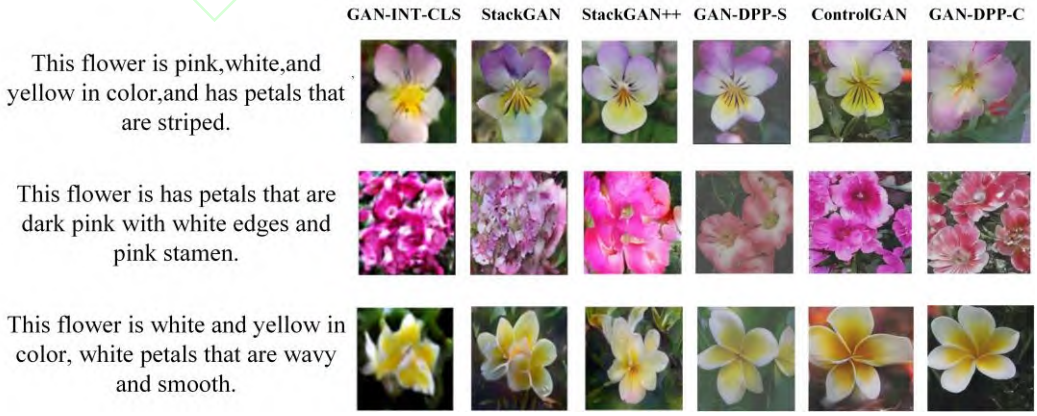

图 6 不同模型在 Oxford-102 测试集上生成图像

Fig.6 Generate Images on the Oxford-102 Test Set in Different Model

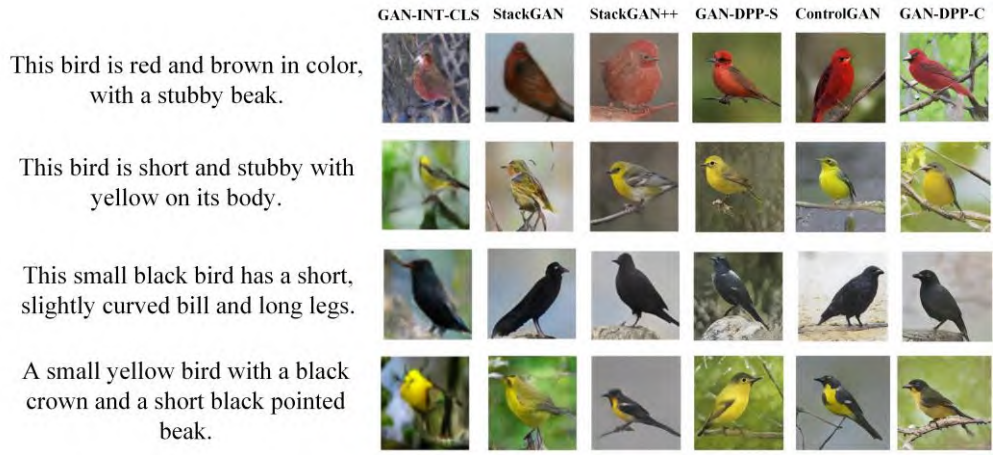

图 7　不同模型在 CUB 测试集上生成图像

Fig.7　Generate Images on the CUB Test Set in Different Models

文本约束的大致花朵颜色与形状，缺乏细节表现。可以观察到 StackGAN 与 ControlGAN 能够生成一些细节纹理，但图片模糊且有失真现象。相比之下 GAN-DPP 模型的图像在花朵颜色与细节表现上更加清晰真实。通过观察第二、三个文本描述对应的图片，可以发现 GAN-DPP-C 生成的图片相比 ControlGAN 色彩更加柔和自然，凭借人的主观感知也难以分辨图像真假。

图 7 展示了各模型在 CUB 数据集上的生成图像对比，通过观察可知，GAN-CLS-INT 模型的图像结果只能够模糊地分辨出背景与对象。StackGAN++在图像分辨率和对象细节表征方面有了提高，但是生成效果仍存在一定缺陷，如图像局部模糊和细节的失真。可以发现 GAN-DPP 模型生成的鸟类身体结构更加合理，并能清楚分辨出嘴和眼睛等重要部分。GAN-DPP-S 相比于 StackGAN++、GAN-DPP-C 相比于 ControlGAN，整只鸟的整个鸟的体态生成真实性有较大提升，并且在符合语义描述的同时，背景内容也更加清晰丰富。

图 8 与图 9 将模型的生成结果进行多样性的对比，针对指定文本观察模型所生成多个图像的表现效果。图 8 为 Oxford-102 数据集上 GAN-DPP-C 模型的生成结果。可以看出图像中花朵形状真实合理，并且对象与背景分明，花瓣花蕊等细节清晰，在真实性方面有很好的表现。图像在多样性方面同样有很大提升，不再局限于一种模式。对于第三个文本，符合语义的同时，生成图像不止存在仅有一朵花在图像中央的模式，同时成功生成了多朵花在同一画面中的情况。对于第四、五个文本所对应的图像，花瓣和花蕊形状、花朵位置与明暗度等特征也呈现了多样的表现。能够观察到，GAN-DPP 使生成模型针对有限文本挖掘出更丰富的图像模式，生成符合约束且更多样的高质量图像。

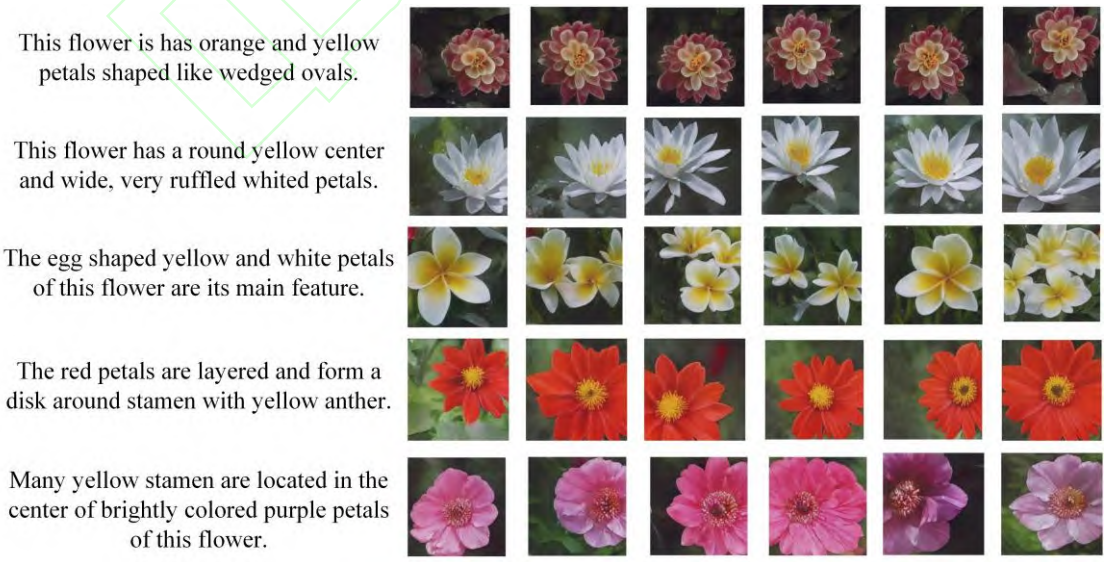

图 8　GAN-DPP-C 模型在 Oxford-102 数据集上的生成图像

Fig.8　Generated Images of the GAN-DPP-C Model on the Oxford-102 Dataset

图 9 为在 CUB 数据集上基于同一文本生成图像的一些示例，将 GAN-DPP-S 模型与 StackGAN++、MSGAN 模型进行对比，观察生成结果的多样性表现。针对同样的文本，StackGAN++模型的生成结果聚集在单一的模式下，图像中鸟类颜色与姿态几乎没有变化，相比之下 MSGAN 与 GAN-DPP-S 生成的鸟类能够有更多的表现形式。并且 GAN-DPP-S 与 MSGAN 相比能够明显地看出，鸟的方向角度、颜色分布等特点表现更加多样，同时图像背景内容也得到进一步丰富。由图 8 图 9 可以看出，GAN-DPP 的模型所生成的样本针对同一文本呈现了变化更加丰富的表征，图像在多样性方面有了明显的提高。

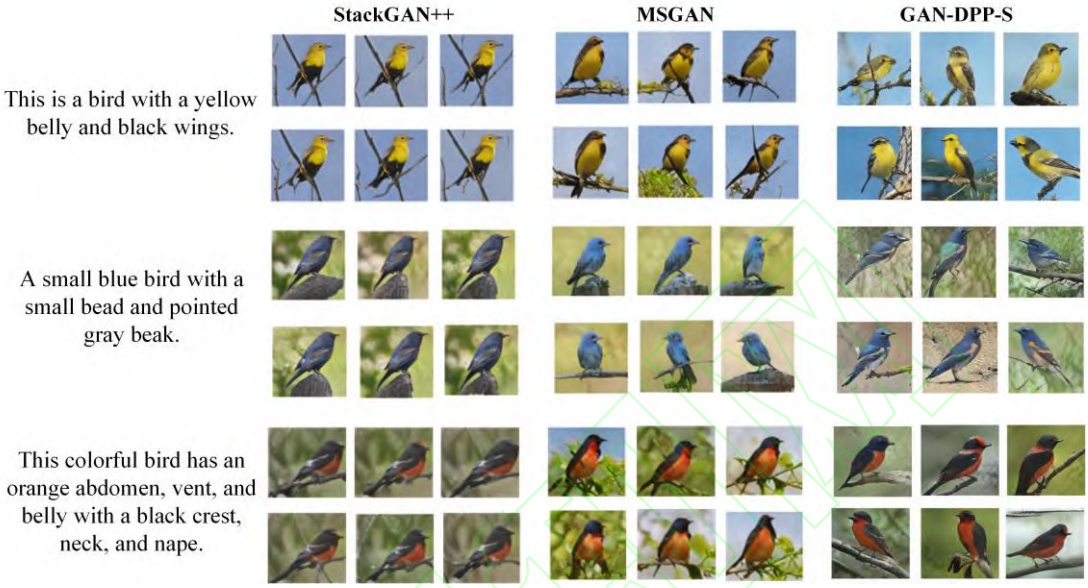

图 9　三种模型在 CUB 数据集上多样性对比

Fig.9　Comparison of the Diversity of the Three Models on the CUB Dataset

3　总 结

本文针对文本生成图像任务，提出了一种提升模型生成图像多样性的方法，即通过引入行列式点过程采样，鼓励生成器生成表现模式更加丰富的图像。通过构建行列式点过程核矩阵，对真实数据和合成数据的多样性进行建模，增加了惩罚项，通过计算 DPP 核矩阵特征值与特征向量之间的损失，训练生成器在生成图像的多样性方面向真实数据分布的多样性靠近，使生成样本在满足生成真实性与条件约束的同时，在多样性方面得到进一步提升。经过定量与定性评估能够看出，对于引入行列式点过程的模型，其生成结果在质量与多样性方面均有更好的表现。但实验仍存在一些不足，对于一些类别的生成样本，存在生成对象扭曲，细节语义有偏差，真实度不够的情况。

参 考 文 献

[1] Wang M, Ai T, Yan X, et al. Grid Pattern Recognition in Road Networks Based on Graph Convolution Network Model[J]. *Geomatics and Information Science of Wuhan University*, 2020, 45(12): 1960-1969

[2] Zheng C X, Cham T J, Cai J F. Pluralistic image completion[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: 1438-1447

[3] Karnewar A, Wang O.MSG-GAN: Multi-Scale Gradient GAN for Stable Image Synthesis[OL]. https://arxiv.org/abs/1903.06048, 2019

[4] Huang Ruobing, Jia Yonghong. Face Swapping Using Convolutional Neural Network and Tiny Facet Primitive[J]. *Geomatics and Information Science of Wuhan University*, 2021, 46(3): 335-340 (黄若冰, 贾永红. 利用卷积神经网络和小面元进行人脸图像替换[J]. 武汉大学学报・信息科学版, 2021, 46(3): 335-340)

[5] Li Y T, Gan Z, Shen Y L, et al. StoryGAN: A sequential conditional GAN for story visualization[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: 6322-6331

[6] Xu K, Ba J L, Kiros R, et al. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention[C]//Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37.2015: 2048-2057

[7] Wei Y C, Zhao Y, Lu C Y, et al. Cross-Modal Retrieval with CNN Visual Features: A New Baseline[J]. *IEEE Transactions on Cybernetics*, 2017, 47(2): 449-460

[8] Goldberg Y. Neural network methods for natural language processing[M]. [San Rafael]: Morgan & Claypool Publishers, [2017]

[9] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative Adversarial Networks[J]. *Communications of the ACM*, 2020, 63(11): 139-144

[10] Mirza M, Osindero S. Conditional Generative Adversarial Nets[OL]. https://arxiv.org/abs/1411.1784, 2014

[11] Odena A, Olah C, Shlens J. Conditional Image Synthesis with Auxiliary Classifier GANs[C].//The 34th International Conference on Machine Learning. Sydney, Australia, 2017

[12] Reed S, Akata Z, Yan X C, et al. Generative Adversarial Text to Image Synthesis[C]//Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48.2016: 1060-1069

[13] Isola P, Zhu J Y, Zhou T H, et al. Image-to-image translation with conditional adversarial networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: 5967-5976

[14] Lu Chuanwei, Sun Qun, Zhao Yunpeng, et al. A Road Extraction Method Based on Conditional Generative Adversarial Nets[J]. *Geomatics and Information Science of Wuhan University*, 2021, 46(6): 807-815 (陆川伟, 孙群, 赵云鹏, 等. 一种基于条件生成式对抗网络的道路提取方法[J]. 武汉大学学报·信息科学版, 2021, 46(6): 807-815)

[15] Nilsback M E, Zisserman A. Automated flower classification over a large number of classes[C]//2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing. Bhubaneswar, India.: 722-729

[16] Wah C, Branson S, Welinder P, et al. The Caltech-UCSD Birds-200-2011 Dataset[J]. *California Institute of Technology*, 2011, 7(1) : 1-8

[17] Zhang H, Xu T, Li H S, et al. StackGAN: text to photo-realistic image synthesis with stacked generative adversarial networks[C]//2017 IEEE International Conference on Computer Vision. Venice, Italy: 5908-5916

[18] Zhang H, Xu T, Li H S, et al. StackGAN: Realistic Image Synthesis with Stacked Generative Adversarial Networks[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, 41(8): 1947-1962

[19] Mao Q, Lee H Y, Tseng H Y, et al. Mode seeking generative adversarial networks for diverse image synthesis[C]//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, CA, USA: 1429-1437

[20] Srivastava A, Valkov L, Russell C, et al. VEEGAN: Reducing Mode Collapse in GANs Using Implicit Variational Learning[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017: 3310-3320

[21] Xu T, Zhang P C, Huang Q Y, et al. AttnGAN: fine-grained text to image generation with attentional generative adversarial networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: 1316-1324

[22] Li B W, Qi X J, Lukasiewicz T, et al. Torr. Controllable Text-to-Image Generation[C]//The International Conference on Neural Information Processing Systems. Vancouver, Canada, 2019

[23] Borodin A. Determinantal Point Processes [OL]. https://arxiv.org/abs/0911.1153, 2009

[24] Reed S, Akata Z, Lee H, et al. Learning deep representations of fine-grained visual descriptions[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: 49-58

[25] Zhou J, Xu W. End-to-end learning of semantic role labeling using recurrent neural networks[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Beijing, China. 2015

[26] Macchi O. The Coincidence Approach to Stochastic Point Processes[J]. *Advances in Applied Probability*, 1975, 7(1): 83-122

[27] Hough J B, Krishnapur M, Peres Y, et al. Determinantal Processes and Independence[J]. *Probability Surveys*, 2006, 3(1): 206-229

[28] Gong B Q, Chao W L, Grauman K, et al. Diverse Sequential Subset Selection for Supervised Video Summarization[C]//Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2.2014: 2069-2077

[29] Elfeki M, Couprie C, Riviere M, et al. GDPP: Learning Diverse Generations Using Determinantal Point Process[OL]. https://arxiv.org/abs/1812.00068v1, 2018

[30] Kulesza A, Taskar B. Structured Determinantal Point Processes[C]//Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 1.2010: 1171-1179

[31] Salimans T, Goodfellow I, Zaremba W, et al. Improved Techniques for Training GANs[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems. 2016: 2234-2242

[32] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: 2818-2826

[33] Heusel M, Ramsauer H, Unterthiner T, et al. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium[C]//Proceedings of the 31st International Conference on

Neural Information Processing Systems. 2017: 6629-6640

[34] Zhang Z Z, Xie Y P, Yang L. Photographic text-to-image synthesis with a hierarchically-nested adversarial network[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: 6199-6208

Determinant Point Process Sampling Method for Text-to-Image Generation

Li Xiaolin[1,2] *Li Gang*[1,2] *Zhang Enqi*[1,2] *Gu Guanghua*[1,2]

1. Department of Information Science and Engineering, Yanshan University, Qinhuangdao 066004

2. Hebei Key Laboratory of Information Transmission and Signal Processing, Yanshan University, Qinhuangdao 066004

Abstract: Objectives: In recent years, a great breakthrough has been made in the text generation image problem based on Generative Adversarial Networks (GAN). It can generate corresponding images based on the semantic information of the text, and has great application value. However, the current generated image results usually lack specific texture details, and often have problems such as collapsed modes and lack of diversity. **Methods:** This paper proposes a Determinant Point Process for Generative Adversarial Networks (GAN-DPP) to improve the quality of the generated samples, and uses two baseline models, StackGAN++ and ControlGAN, to implement GAN-DPP. During the training, it uses Determinantal Point Process kernel to model the diversity of real data and synthetic data and encourages the generator to generate diversity data similar to the real data through penalty loss. It improves the clarity and diversity of generated samples, and reduces problems such as mode collapse. No extra calculations were added during training. **Results:** This paper compares the generated results through indicators. For the Inception Score score, a high value indicates that the image clarity and diversity have improved. On the Oxford-102 dataset, the score of GAN-DPP-S is increased by 3.1% compared with StackGAN++, and the score of GAN-DPP-C is 3.4% higher than that of ControlGAN. For the CUB dataset, the score of GAN-DPP-S increased by 8.2%, and the score of GAN-DPP-C increased by 1.9%. For the Fréchet Inception Distance score, the lower the value, the better the quality of image generation. On the Oxford-102 dataset, the score of GAN-DPP-S is reduced by 11.1%, and the score of GAN-DPP-C is reduced by 11.2%. For the CUB dataset, the score of GAN-DPP-S is reduced by 6.4%, and the score of GAN-DPP-C is reduced by 3.1%. **Conclusions:** The qualitative and quantitative comparative experiments prove that the proposed GAN-DPP method improves the performance of the generative confrontation network model. The image texture details generated by the model are more abundant, and the diversity is significantly improved.

Key words: Generative Adversarial Networks; Text-to-image Synthesis; Determinantal Point Process; Mode Collapse; Diversity

First author: Li Xiaolin, postgraduate, specializes in Generative Adversarial Networks, text-to-image generation. E-mail: imlixlin@163.com

Corresponding author: Gu Guanghua, PhD, professor. E-mail: guguanghua@ysu.edu.cn

Foundation support: The National Natural Science Foundation of China(No.62072394); the National Natural Science Foundation of Hebei province(F2021203019)