基于扩散模型的文本连续性插画生成

徐任政[1] 姚剑敏[1,2] 陈恩果[1] 严 群[1]
XU Renzheng YAO Jianmin CHEN En'guo YAN Qun

摘 要 基于扩散模型的生成能力，提出了一种基于扩散模型的文本连续性插画的生成方法。其目标是根据给定的文本输入，生成连续的插画，以便更好地传达文本的意思。在扩散模型的基础上，加入了 LSTM 神经网络模型，使扩散模型在原本只能根据一句或者一段文本来生成一张图片的基础上加以改进，之后可以通过一段或多段文本，通过 LSTM 进行预处理，再生成一系列连续性的图片。所提方法主要是利用 LSTM 神经网络模型在序列建模方面的优势，它能很好地捕捉序列数据中的长期依赖关系，从而更好地理解和建模序列中的上下文信息。将预处理好的源文本送入生成模型，通过生成模型的生成能力，最后输出一系列连续的图片。

关键词 扩散模型；深度学习；插画生成；连续性；LSTM 神经网络模型

doi：10.3969/j.issn.1672-9528.2024.04.005

0 引言

插画是一种常常运用在书籍、杂志、漫画、广告等媒体中的艺术形式。人们通常是通过手动作画、绘图软件作画或者是通过其他数字工具来制作，目的是补充和解释文字内容，方便读者加深理解。插画可以通过创造性的图像来传达信息、故事和情感，在设计和传媒等领域有着广泛的应用。

然而，由于时间和精力等因素的制约，对于一些大篇幅的文章或者书籍，一幅幅地去绘制插画显然是较为困难的，即使通过一些单纯的生成对抗网络去实现，也很难达到预期的效果。因此，本文所提出的连续性插画生成方法，是解决这一问题的重要方式之一。

近年来，基于深度学习的生成式建模方法已经得到广泛应用。主要的生成式建模方法有生成对抗网络（GAN）、变分自编码器（VAE）、扩散模型（diffusion models）等[1]。其中，扩散模型得益于其自身生成样本的强大能力，已被广泛应用于高质量图像生成[2]、自然语言处理和生物信息学等多个领域。Jascha Sohl-Dickstein 等人[3]第一次提出了扩散概率模型（diffusion probabilistic model，简称扩散模型），其基本思想是通过不断迭代来系统地、缓慢地破坏数据的分布结构，再从逆向过程中恢复数据结构，从而构建一个灵活且易于处理的数据生成模型。不仅如此，用于自然语言处理的 LSTM 神经网络，同样成为计算机视觉方面的研究热点，在“看图说话”“文本情感分析”[4]等领域得到广泛应用[5-6]。

本文主要研究如何利用扩散模型的生成能力来实现文本连续性插画生成。过去的生成模型研究主要注重文本到图像的转换，使得文本的描述能够被转化为静态的图像呈现。然而，这种呈现方式往往无法捕捉到文本的连续性和动态特点，限制了信息传达的效果。为了解决这一问题，本文提出了基于扩散模型的文本连续性插画生成。本文以扩散模型的生成能力为基础，结合 LSTM 神经网络模型[7]，实现连续性插画生成。

1 模型构建

1.1 LSTM 神经网络模型

LSTM 神经网络模型也被称为长短期记忆网络。它能解决 RNN（循环神经网络）模型中的梯度消失和梯度爆炸问题，以及长期依赖关系的建模能力不足的问题，它在自然语言处理（NLP）任务中得到广泛应用。RNN 是想把所有信息都记住，不管是有用还是没有用的信息。而 LSTM 是通过设计一个记忆细胞，让它具备选择性记忆的能力，可以选择记忆重要的信息，过滤掉噪声信息，减轻记忆负担。

LSTM 单元主要由三个门控单元组成：遗忘门（forget gate）、输入门（input gate）以及输出门（output gate）。每个门都包含了一个 Sigmoid 激活函数，用于控制门的打开和关闭程度。此外，LSTM 单元还具有一个记忆单元（cell state），用于存储和传递历史信息。

遗忘门：对于某个时刻来说，之前的一些“信息”可能

1. 福州大学 福建福州 350108
2. 晋江市博感电子科技有限公司 福建晋江 362200
[基金项目] 国家重点研发计划（2022YFB3603503），福建省技术攻关重点项目（2023G007）

已经“过时”了。为了不让这些已经“过时”的信息影响到现在的状态，人们会选择去“遗忘”掉这些“过时”的信息。可通过一个 sigmoid 神经层来过滤掉这些“过时”的信息，其公式为：

$$f_t = \sigma(W_f \cdot [C_{t-1}, X_t] + b_f) \quad (1)$$

记忆门：用来控制在 t 时刻的数据并入单元状态中的控制单位。首先用 tanh 函数提取出向量中的有效信息，接着通过 Sigmoid 函数来控制和筛选输入到单元状态中的信息，其公式为：

$$i_t = \sigma(W_i \cdot [C_{t-1}, X_t] + b_i) \quad (2)$$

$$C_t = \tan(W_c \cdot [h_{t-1}, X_t] + b_c) \quad (3)$$

输出门：LSTM 单元用于计算当前时刻的输出值的神经层。将经过 tanh 函数处理后的数据，再通过 Sigmoid 函数处理，整合后通过向量点乘计算即可得到 LSTM 在 t 时刻的输出。

为了使 LSTM 神经网络模型更加直观和容易理解，图 1 展示了该神经网络模型的内部结构。

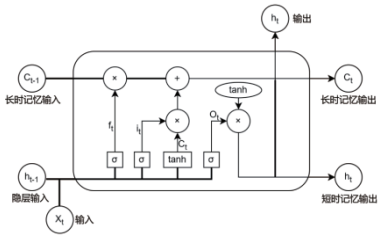

图 1 LSTM 网络结构图

1.2 LSTM 在图像生成中的应用

LSTM 是一种循环神经网络（RNN）的变体，主要应用于序列数据的建模和预测。尽管 LSTM 最开始的初衷是用于自然语言处理，但是随着不断地探索，它也逐渐被应用于其他的领域，如图像描述生成、图像翻译生成等。

在图像生成中，LSTM 可以用于生成图像的描述或者从给定的文本描述中生成图像。LSTM 可以将输入的文本序列编码为一个固定长度的向量表示，然后使用解码器将该向量输入到神经网络中，生成与文本描述相匹配的图像，它本身并不直接生成图像像素级别的内容。LSTM 可以在图像生成任务中参与与图像相关的自然语言描述部分，从而加强图像的语义理解能力和应用场景的拓展。

1.3 扩散模型（Diffusion Models）

如图 2 所示，是扩散模型的基本结构。扩散模型可分为扩散过程和逆扩散过程两部分 [8-9]，扩散过程就是不断对原始数据加入高斯噪声，让原始数据的分布转换为一个简单的标准高斯分布 [10]。逆扩散是一个去噪的过程，从标准高斯分布中进行采样，每一步去除一个很小的高斯噪声，逐步贴近真实数据分布，进而得到其真实数据分布中的样本，从而达到生成数据的目的 [11-12]。扩散模型的特点是可以生成高质量的图片，并能够对生成的图片进行控制。通过在生成过程中引入不同的条件，可以实现多样性和个性化。它还被应用于数据修复、降噪、图像增强等许多领域。

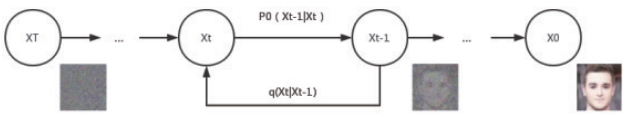

图 2 扩散模型的基本结构图

前向过程：它是加噪的过程 [13]。前向过程中，图像 x_t 只和上一个时刻的 x_{t-1} 有关，这个过程可以视为马尔可夫过程，它满足：

$$q(x_t|x_{t-1}) = N\left(x_t; \sqrt{(1-\beta_t)}x_{t-1}, \beta_t I\right) \quad (4)$$

$$q(x_1, x_2, \cdots, x_T|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \quad (5)$$

式中：不同 t 的 β_t 是预先定义好的逐渐衰减的，可以是 Linear、Cosine 等，满足：

$$0 < \beta_1 < \beta_2 < \cdots < \beta_T < 1 \quad (6)$$

经过计算可得：

$$x_t = \sqrt{(1-\beta_t)}x_{t-1} + \sqrt{\beta_t}z_{t-1} \quad (7)$$

式中：$z_{t-1} \sim N(0,I)$，$\alpha_t = 1-\beta_t$。

经过推导，可以得到 x_t 与 x_0 的关系：

$$q(x_t|x_0) = N\left(x_t; \sqrt{\alpha}x_0, (1-\overline{q}_t)I\right) \quad (8)$$

逆向过程：它是去噪的过程，使用神经网络 $p_\theta = (x_{t-1}|x_t)$ 拟合逆过程 $q(x_{t-1}|x_t)$。可以通过：

$$q(x_t|x_0) = N\left(x_t; \sqrt{\alpha}x_0, (1-\overline{\alpha}_t)I\right) \quad (9)$$

推导出：

$$p_\theta(x_{t-1}, x_t) = N\left(x_{t-1}|\mu_\theta(x_t, t), \sum_\theta(x_t, t)\right) \quad (10)$$

1.4 LSTM 与扩散模型的结合

LSTM（长短期记忆网络）和扩散模型是两种不同的神经网络模型，它们分别适用于不同的任务和领域。然而，将它们结合起来可以带来一些独特的优势，特别是在连续性插画生成任务中。LSTM 在序列建模方面表现出色，它可以学习到序列数据中的上下文信息和时间依赖关系。扩散模型则在图像处理中具有擅长的特征提取和表示学习能力。将 LSTM 和扩散模型结合，可以通过 LSTM 学习到的上下文信息指导扩散模型对图像特征进行更准确的提取和表示学习。通过 LSTM 与扩散模型的结合，可以提高模型的生成能力、长期依赖关系性建模能力以及对重要特征的关注能力。这种结合可以应用于各种序列生成任务，例如自然语言处理、图像生成等。

同样，LSTM 与扩散模型结合在插画的连续性生成任务中也拥有出色的表现。LSTM 能够学习到插画序列的连贯性

和结构性，而扩散模型则能够学习到插画中的局部特征和细节。通过将两者结合，可以在生成插画时既保持整体的连贯性，又注重局部的细节和质感，从而生成更加真实且富有表现力的连续性插画。

1.5 网络结构设计

将 LSTM 和扩散模型结合，可以设计出一种新颖的网络结构，实现更好的插画生成效果。如图 3 所示，本文通过使用文本编码器和图像编码器，分别对输入的文本与图像进行编码，得到文本与图像的特征表示。然后，将得到的文本特征和图像特征进行连接，形成联合的特征表示。接着，将联合特征输入到 LSTM 解码器中，解码器按照输入的特征生成一系列的特征向量。最后，通过残差连接的方式将这些特征向量输入到扩散模型中，由扩散模型来生成图像。

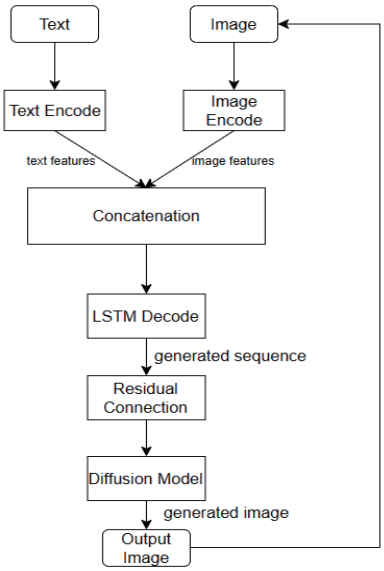

图 3 改进网络结构图

为了保持连续性和一致性，本文在扩散模型的特征输入上应用残差连接，以获得更好的生成图像效果。之后，将生成的图像作为下一个时间的输入，继续结合文本生成下一张图像，直到所有插画都生成完毕。这个网络结构充分利用了 LSTM 在序列建模方面的优势以及扩散模型在生成方面的优势，巧妙地将两种结构的优势互相结合，这使得模型能够更好地表达文本的含义和情感，因此能够在连续性的插画生成任务中达到更加理想的效果。

2 实验与结果与分析

2.1 数据集

本次实验所采用的数据集是 Flickr 数据集。Flickr 数据集是一个公开可用并且被广泛使用的图像数据集，由雅虎公司创建并维护。Flickr 数据集提供了丰富的图像信息和标签信息，这些都可以用于训练和评估与图像相关的模型。该数据集的多样性及其庞大的规模，使它成为许多研究工作的基准数据集之一。因此，它也是计算机视觉研究和开发的重要资源。研究人员和开发者可以利用该数据集来进行图像分类、目标检测、图像生成、图像描述和视觉推理等任务的训练和评估。

2.2 实验条件

实验在 Ubuntu、Window10 环境下进行，使用的编程环境为 Pycharm，编程语言使用的是 Python3.7，深度学习框架使用的是 Pytorch 和 CUDA。硬件设备采用 i7-6700CPU+Tesla P40 显卡服务器，运行内存 16 GB（RAM）。

2.3 实验结果及分析

两组图片根据相同的 text 来生成：“Mike and Jon went to explore the forest together. They passed by a small stream and saw many small fish inside.Mike curiously crouched down and cautiously reached out to get into the water, trying to catch one of the small fish. However, the little fish was agile and could always quickly dodge before his fingers touched them, transforming into a lightning like figure.Jon couldn’t help but laugh as he watched Mike’s movements. Mike also laughed along, shaking his head and sitting on the big stone by the shore.”

如图 4 所示，是根据文本直接由扩散模型生成的图片。而图 5 是由本文构建的后的网络架构生成的图片。每张图片分别对应一句 text 文本，从左上到右下依次是第一张到最后一张。通过对比两组图片，可以看出，扩散模型的生成能力天马行空，它并未与其他图片的特征相关联，只和输入的文本、关键词有关。它生成的每一张图片都有许多的可能性。从图 4 中，也很难看出图片的连续性。而通过改进后的网络结构生成的图片，如图 5 所示，可以更加明显地看出它们在风格上相比于图 4 更加接近，人物以及背景的转换更加连贯和一致。

图 4 扩散模型生成图

图 5 LSTM 与扩散模型结合生成图

虽然实验生成的连续性插画达到预期效果，但是仔细观察，插画的质量还有待提升。例如，插画中人物的着装、表情、动作等一系列变换，还是不能根据文本的实际含义完成平滑的转换。并且，插画生成的风格也需要根据需求来训练不同的数据集，这会造成在不同的任务中，可能需要重新去训练的问题。在实际应用中，可以根据具体问题的特点选择合适的方法去解决，例如用不同的数据集训练、改变网络结构等。

总体来说，改进后的网络结构，相比于直接使用扩散模型，无疑是更适合用于连续性插画的生成。从生成的结果对比来看，它在连续性生成领域具有明显的优势，并且也拥有更加广泛的应用前景。

3 结语

本文研究并探索了将 LSTM 神经网络和扩散模型结合并用以生成连续性插画的方法。通过使用 Flickr 数据集进行训练，目标是在图像生成任务中实现平滑的图像转换、语义一致性和高逼真度。

从实验结果来看，LSTM 与扩散模型能够有效结合，从而生成连续性插画。生成的连续性插画效果达到预期。这种结合的方法为插画生成任务提供了一个较好的框架，可以在许多领域中发挥重要作用，提升工作效率。

尽管本次实验取得了一些令人鼓舞的结果，但也存在局限性。受到训练数据分布的限制，生成的某些复杂场景或对象可能效果并不是很理想。对于不同的任务要求以及不同的数据集，为了提高生成插画的效果，可以对结构进行进一步的优化和调整。

总而言之，本文的研究为 LSTM 与扩散模型结合生成连续性图片提供了一种有前景的方法，能为相关研究和应用提供有价值的参考和启示，对于计算机视觉和图像生成领域的进一步发展具有重要意义。

参考文献：

[1] 杨光锴 . 基于扩散模型的指纹图像生成方法 [J]. 河北省科学院学报 ,2023(1):13-18+66.

[2]DOSOVITSKIY A, SPRINGENBERG J T, BROX T.Learning to generate chairs with convolutional neural networks[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).Piscataway: IEEE,2015:1538-1546.

[3]JASCHA S D, ERIC A W, NIRU M, et al.Deep unsupervised learning using nonequilibrium thermodynamics[C]// 32nd International Conference on Machine Learning, volume 3 of 3. New York:Curran Associates, Inc.,2016:768-784.

[4] 王志平 , 郑宝友 , 刘仪伟 . 一种改进的 LSTM 模型在图像标题生成中的应用 [J]. 计算机与现代化 ,2020(4):37-41

[5] 冯兴杰 , 张志伟 , 史金钏 . 基于卷积神经网络和注意力模型的文本情感分析 [J]. 计算机应用研究 ,2018,35(5):1434-1436.

[6]KELVIN X, JIMMY B, RYAN K, et al.Show, attend and tell: neural image caption generation with visual attention[C]// 32nd International Conference on Machine Learning,volume 3 of 3.New York:Curran Associates, Inc.,2016:477-499.

[7] 王燕萍 , 吕磊 , 苏志龙 , 等 . 基于深度学习的高质量图像生成方法综述 [J]. 激光杂志 ,2023(6):7-12.

[8]IAN J G, JEAN P A, MEHDI M.Generative adversarial nets[EB/OL].(2014-06-10).https://arxiv.org/pdf/1406.2661.pdf.

[9]RONNEBERGER O, FISCHER P, BROX T.U-Net: convolutional networks for biomedical image segmentation[EB/OL].(2015-05-18).https://arxiv.org/abs/1505.04597.

[10] 卢玲 , 杨武 , 王远伦 , 等 . 结合注意力机制的长文本分类方法 [J]. 计算机应用 ,2018,38(5):1272-1277.

[11]YANG L, ZHANG Z L, SONG Y, et al.Diffusion models:a comprehensive survey of methods and applications[J].ACM computing surveys,2022,56:1-39.

[12]GREGOR K, DANIHELKA I, GRAVES A, et al. DRAW: a recurrent neural network for image generation[C]//32nd International Conference on Machine Learning,volume 2 of 3.New York: Curran Associates, Inc., 2016:906-915.

[13]CAO H Q, TAN C, GAO Z Y, et al.A survey on generative diffusion model[J].IEEE transactions on knowledge and data engineering,2022,14:1-20.

【作者简介】

徐任政（2000—），男，福建莆田人，硕士研究生，研究方向：深度学习、图像处理等。

姚剑敏（1978—），男，福建莆田人，博士，副研究员，研究方向：人工智能、图像处理、计算机视觉等。

（收稿日期：2024-01-15）