从像素卷积神经网络隐空间中生成新桥型的尝试

张洪俊

（万世先行数智交通科技有限公司，南京 210016）

摘要：尝试利用生成式人工智能技术生成新桥型。采用对称结构的三跨梁式桥、拱式桥、斜拉桥、悬索桥图像数据集，基于 Python 编程语言、TensorFlow 及 Keras 深度学习平台框架，构建和训练像素卷积神经网络。模型能够捕捉图像的统计结构，在给出前面像素的条件下，能够计算出下一个像素的概率分布。从得到的隐空间采样，能够生成不同于训练数据集的新桥型。像素卷积神经网络能够在人类原创桥型的基础上，将不同结构部件进行有机搭配，创造生成新桥型，一定程度上具有类似人类的原创能力。自回归模型不能明白序列所表达的含义，而多模态模型将回归和自回归模型结合，从而能够理解序列。多模态模型应该是未来实现通用人工智能的路径。

关键词：生成式人工智能；桥型创新；像素卷积神经网络；隐空间；自回归模型；深度学习

中图分类号：U448.2；TP181　　**文献标志码：**A

An attempt to generate new bridge types from latent space of PixelCNN

ZHANG Hong-jun

Wanshi Antecedence Digital Intelligence Traffic Technology Co., Ltd, Nanjing, 210016, China

Abstract: Try to generate new bridge types using generative artificial intelligence technology. Using symmetric structured image dataset of three-span beam bridge, arch bridge, cable-stayed bridge and suspension bridge , based on Python programming language, TensorFlow and Keras deep learning platform framework , PixelCNN is constructed and trained. The model can capture the statistical structure of the images and calculate the probability distribution of the next pixel when the previous pixels are given. From the obtained latent space sampling, new bridge types different from the training dataset can be generated. PixelCNN can organically combine different structural components on the basis of human original bridge types, creating new bridge types that have a certain degree of human original ability. Autoregressive models cannot understand the meaning of the sequence, while multimodal models combine regression and autoregressive models to understand the sequence. Multimodal models should be the way to achieve Artificial General Intelligence in the future.

Keywords: generative artificial intelligence; bridge-type innovation; PixelCNN; latent space; autoregressive model; deep learning

0　引言

在人类历史长河中，许多新技术的出现如材料、力学、计算机等，对桥梁工程产生了巨大甚至革命性的影响。17 世纪以前，桥梁一般采用木材、石料建造，受制于材料力学性能，最大跨径仅几十米。18 世纪工业革命后，铁、钢、混凝土的生产，为桥梁提供了新的建造材料，很快桥梁最大跨径就可达到几百米。在挠度理论指导下 1937 年建成的金门大桥跨径达到 1280 米。20 世纪 60 年代诞生的基于计算机的有限元分析法，能够精确地分析复杂结构的力学行为，从而让现代斜拉桥桥型得以实现。目前桥梁最大跨径已突破 2000 米，如果没有先进的材料、力学理论和计算手段，这根本无法实现。

当前人工智能技术，为土木工程发展提供了新的动力，相信智能建造时代即将来临[1]，而本文的桥型创新仅是其应用的沧海一粟。可以想象不远的未来：虚拟助手采用混合现实的方式，向人类工程师汇报它的创作；智能机器在别墅施工工地上忙碌着，人类则喝着咖啡憧憬建成入住的那天。

作者之前论文[2-3]训练变分自编码器（Variational Autoencoder，VAE）和生成对抗网络（Generative Adversarial Network，GAN），成功地生成了与数据集完全不同的新桥型，得出了它们能够协助桥梁设计师进行桥型创新的结论。在众多的生成式人工智能技术中，自回归模型（Autoregressive Models）、基于能量的模型（Energy-Based Models）和扩散模型（Diffusion Models）等，同样可以应用于桥型创新。

像素卷积神经网络(Pixel Convolutional Neural Network, PixelCNN)将图像设想为像素序列，根据之前像素来计算下一个像素的概率分布，逐像素生成图像。本文采用 PixelCNN，基于之前同样的数据集，进一步尝试桥型创新（本文数据集与源代码开源地址 https://github.com/QQ583304953/Bridge-PixelCNN）。

1　PixelCNN 简介

1.1 自回归模型概述

回归模型是根据自变量(X)来估计因变量(Y)，例如根据基本面分析来预测股市走势（利好利空概念）。

自回归模型是根据自身过去的值来估计当前值，例如仅根据股市自身的历史数据来预测当前股市走势（K 线图技术分析）。

自回归模型可以揭示时间序列数据的内在规律和趋势，常见的有 LSTM、PixelCNN 和 GPT 等模型。它打败过人类棋手，在股市中采用趋势跟踪策略可以获利，对地震发生的历史记录进行统计从而能够预测地震，使 ChatGPT 类似人类一样聊天交流。

1.2 LSTM 逐字符生成文本的步骤

以参考文献《python 深度学习》[4]P227 第 8.1 节为例：①先从语料库中提取众多的样本与目标（样本为 60 个字符序列、目标为样本的下一个字符）；②训练 LSTM 语言模型；③然后给出初始文本（种子序列），输入模型，得到下一个字符的概率分布，按照概率值随机采样（如下一个字符是 a 的概率为 0.3，那么会有 30%的概率选择它），逐字符连续生成一整段文本，详见图 1。

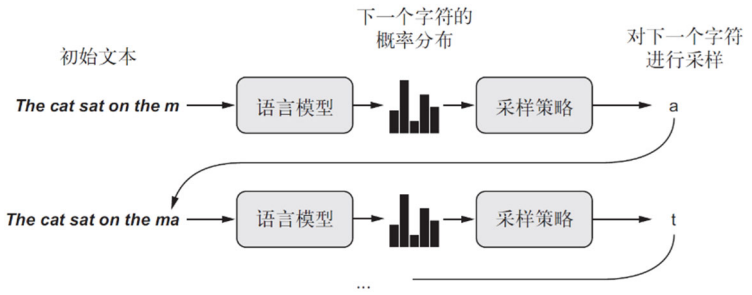

图 1 使用语言模型逐字符生成文本的过程

Fig.1 The process of character-by-character text generation using a language model

1.3 PixelCNN 逐像素生成图像的步骤

过程与 LSTM 逐字符生成文本完全雷同，以灰度图像为例：

（1）将图像设想为像素序列[5-7]，从左到右、从上到下地给图像每个像素标上序号。

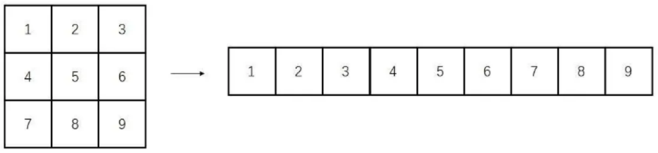

图 2 图像转换为像素序列

Fig.2 The image is converted to a sequence of pixels

（2）数据集：图像一小块区域为样本，目标为该区域的中心像素值。3x3 卷积核时，图 2 的像素序号 1~4 为样本，像素序号 5 为目标，像素序号 6~9 因与任务无关故被屏蔽。每张 192x48 像素的灰度图像相当于 192*48=9216 个训练样本。

（3）训练 PixelCNN。

（4）Conv2D 层的参数 padding="same"，从而使得边角像素也能作为卷积窗口的中心，相当于给出了种子像素序列，然后模型根据前面像素来输出下一个像素的概率分布，按照概率值随机采样，逐像素生成图像。

在生成图像时，先设定空图像，接着模型代码从头至尾运算一遍，算得第一个像素预测值；然后把第一个像素预测值填入空图像，输入模型，得到第二个像素……。一张图片有多少个像素，模型就运行多少遍，单张图片不同像素无法并行计算，所以生成速度很慢。这导致需要很高的算力才能实时监控图像生成效果，模型调试耗时漫长。

1.4 PixelCNN 的隐空间

图像的样本空间各个维度高度相关（如对称图像间隔很远的像素值是相等的），空间维度数目=像素数目，每个维度的离散值数目=像素类别数目。样本空间非常庞大，样本容量=像素类别数目^维度数，这是一个天文数字，空间中大部分样本点没有实际意义。

如何处理空间维度之间的依赖关系、如何从空间中找到有意义的样本，是所有生成式深度学习的难题。直接对样本空间的联合概率分布建模，会因所需样本数量、模型参数量巨大而难以实际操作。

如果采用朴素贝叶斯算法建模（维持原始样本空间的维度数目不变，假设各个维度是完全独立的），那么将会完

全割裂各个维度之间的关联。这导致与实际的情况偏差巨大，在如此高维的隐空间中采样得到令人满意样本的可能性，如同一个猴子不停地敲击键盘就能写出一本小说一般。

与 VAE、GAN 的低维压缩、维度相互独立的隐空间不同，PixelCNN 的隐空间维持原始样本空间的维度数目不变。它通过数据统计，来捕获隐空间相邻维度之间的依赖关系，以条件概率乘积来近似表示联合概率分布。具体为：①图像在原始样本空间中的联合概率分布为 $p(x)=p(x_1, x_2, \cdots x_n)$；②建立一个与原始样本空间维度数目相同的隐空间，设置像素序列（维度的先后次序），假设一个像素的值仅取决于它之前的像素值，即 $p(x_i)= p(x_{i-1}) * p(x_i \mid x_{i-1})$；③如此就可以将联合概率分布近似为条件概率的乘积[5-7]，即 $p(x)= p(x_1)* p(x_2 \mid x_1)* p(x_3 \mid x_2) \cdots p(x_n \mid x_{n-1})$。

显然 PixelCNN 的这种近似操作是不足够精确的，会与实际情况有一定的偏差，但贵在将复杂问题简单化，便于统计学习。这类自回归模型在文本生成和语音生成等领域已经取得了很好的应用成果。

2 从 PixelCNN 隐空间中生成新桥型的尝试

2.1 数据集

采用作者之前论文[2-3]的数据集，即每种桥型两种子类（分别为等截面梁式桥、V 形墩刚构梁式桥、上承式拱式桥、下承式拱式桥、竖琴式斜拉桥、扇式斜拉桥、竖吊杆悬索桥、斜吊杆悬索桥），且均为三跨（梁式桥为 80+140+80m，其它桥型均为 67+166+67m），结构对称。

本模型参数量较大、生成图像速度极慢，故将图片尺寸由 512x128 减少为 192x48 像素（缺点是清晰度降低了许多）。

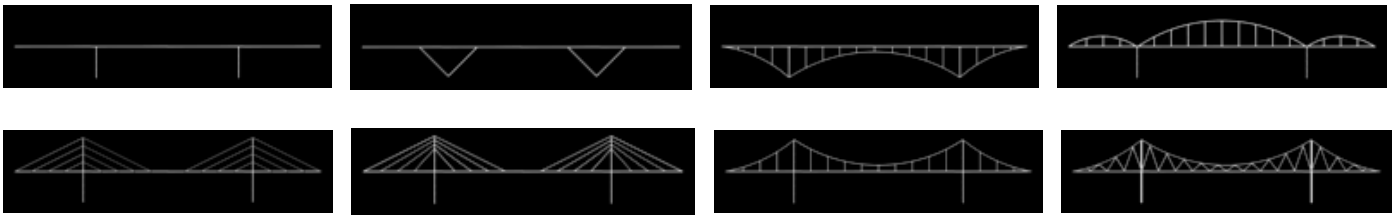

图 3 各桥型立面灰度图

Fig. 3 Grayscale image of each bridge facade

每个子类桥型 1200 张各不相同的图片，整个数据集共 9600 张图片。

2.2 PixelCNN 构建

基于 Python3.10 编程语言、TensorFlow2.10 及 Keras2.10 深度学习平台框架，构建和训练 PixelCNN，总体架构图如下：

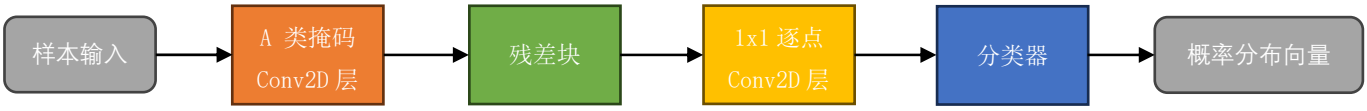

图 4 PixelCNN 总体架构图

Fig. 4 Overall architecture of pixelcnn

（1）A 类掩码 Conv2D 层

为了使每个像素层的输出仅受相关像素之前的像素值的影响，需要屏蔽卷积滤波器窗口的局部，这里通过掩码与过滤器权重矩阵相乘来实现的，以便将目标像素之后的任何像素的值归零[5-7]。

初始屏蔽卷积层不能使用中心像素，因为这正是我们希望网络猜测的像素，即采用 A 类掩码。

后续层（残差块、x1 逐点 Conv2D 层）可以使用中心像素，因为它是根据原始输入图像先前像素的信息计算得出的中间结果，即采用 B 类掩码。

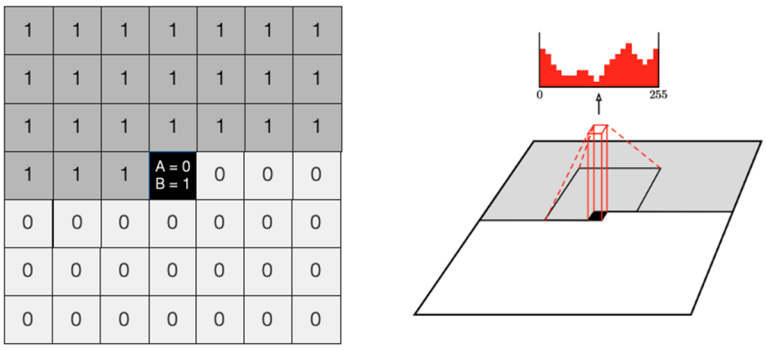

图 5 左 - 卷积过滤器掩码，A 型遮盖中心像素，B 型不遮盖中心像素。右 - 应用于一组像素以预测中心像素值分布的掩码

Fig.5 Left-convolution filter mask, type A covers the center pixel, type B does not cover the center pixel. Right - A mask applied to a set of pixels to predict the distribution of central pixel values

（2）残差块

残差块是一组层，其输出在传递到网络的其余部分之前，与输入相加。换句话说，输入有一条到输出的快速路径，无需经过中间层，这称为跳跃连接。

残差连接可以解决梯度消失和表示瓶颈的问题[4]。

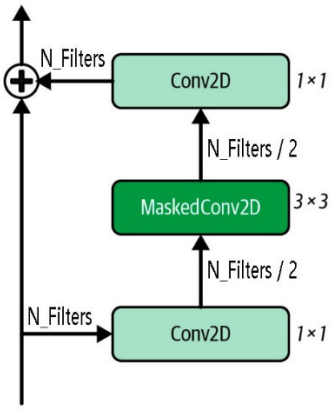

图 6 PixelCNN 中的残差块

Fig.6 A residual block in PixelCNN

（3）1x1 逐点 Conv2D 层

1×1 逐点卷积运算等价于让每个像素向量（多个通道值）经过单个 Dense 层，它计算得到的特征能够将输入张量通道中的信息混合在一起，但不会将跨空间的信息混合在一起（因为它一次只查看一个像素），它有助于区分开通道特征学习和空间特征学习。如果每个通道在跨越空间时是高度自相关的，但不同的通道之间可能并不高度相关，那么这种做法是很合理的[4]。

（4）分类器

输出层是一个 softmax 层，它预测像素的所有可能类别。

2.3 训练

（1）初始版本的 PixelCNN 有个挑战，即模型无法理解像素值 254 与 255 是非常接近的，它必须独立学习每个像素输出值，过多的目标类别导致需要的训练轮数剧增。

假设有 2 个像素样本，像素值目标为[254，254]。PixelCNN 预测的 256 个像素值概率分布为[[0.01，0.06，…，0.8，0.05]，[0.02，0.03，…，0.05，0.8]]，即第一个样本预测的最大概率对应像素值为 254、第二个样本预测的最大概率对应像素值为 255，从人类视觉感受角度而言，预测都很准确。而 sparse_categorical_crossentropy 交叉熵损失为[0.22，3.0]，两个样本损失差别巨大，在模型看来这是预测类别发生了错误。

改进的方法之一是减少类别数目，将 0~255 像素值划分为几个区间（区间就是类别），使每个像素只能取几个类别中的一个。类别数目的降低，模型更容易求解，代价是图像只能由少数几种颜色表达（彩色图像时这个问题会非常突出）。

（2）当概率分布曲线呈现连续光滑时，相邻像素的概率差别会很小，此时相当于模型能够理解像素值 254 与 255 是非常接近的，而不是类别的差异。这就是 PixelCNN++版本[8]的主要改进之处，它采用 logistic 分布计算概率。将 logistic 分布离散为 256 个区间，每个区间依次对应 0~255 像素值。模型输出均值、标准差两个参数，即可完全控制 logistic 曲线的高矮胖瘦，从而精准控制各个区间的概率。

这里直接采用 TensorFlow Probability 库[9]中的 PixelCNN 类（PixelCNN++版本），其拥有优越的性能，直接套用官网 MNIST 实例代码，即可取得初步成果，极大地方便了使用者。微调了 num_resnet、num_hierarchies、num_logistic_mix、receptive_field_dims 四个参数就可以获得较满意的结果。具体参数如下：

```
dist = tfp.distributions.PixelCNN(
    image_shape=(48, 192, 1),
    num_resnet=3, #最高级别块中残差层数目，官网MNIST实例取1，tfp默认5
    num_hierarchies=1, #最高级别块的数目(arXiv:1701.05517v1图2)，官网MNIST实例取2，tfp默认3
#查看源码"pixel_cnn.py": "for i in range(self._num_hierarchies)".每个"i"迭代构建一个最高级别的块
# （在arXiv:1701.05517v1的图2中被标识为"6层序列"，由"num_resnet=5"阶-1层和一个收缩高度/宽度维度的阶-2层组成）
    num_filters=32, #神经元数目，官网MNIST实例取32，tfp默认160
    num_logistic_mix=1, #几个logistic分布组成的混合分布，官网MNIST实例取5，tfp默认10。灰度图像取1可加快损失下降
    receptive_field_dims=(5,7), #tfp默认(3,3)。查看源码"pixel_cnn.py": rows, cols = receptive_field_dims;
#Conv2D( kernel_size=(2 * rows - 1, cols),...)，可见第一个值是可见像素的最大行数、第二个值卷积核窗口列数。
    dropout_p=0.3, #官网MNIST实例取0.3，tfp默认0.5
) #其它参数默认。https://tensorflow.google.cn/probability/api_docs/python/tfp/distributions/PixelCNN
```

图 7 TensorFlow Probability 库中的 PixelCNN 类参数设置

Fig.7 Parameter Settings of pixelcnn class in tensorFlow probability library

受硬件条件制约，不可能充分优化，所以只能边测试参数边采样与筛选，生成桥型图像来自多个事先保存的权重模型。

2.4 隐空间采样探索新桥型

随机采样，然后人工基于工程结构思维筛选，得到了与训练集完全不同的 12 种技术可行的新桥型（图 8，白底黑字）。

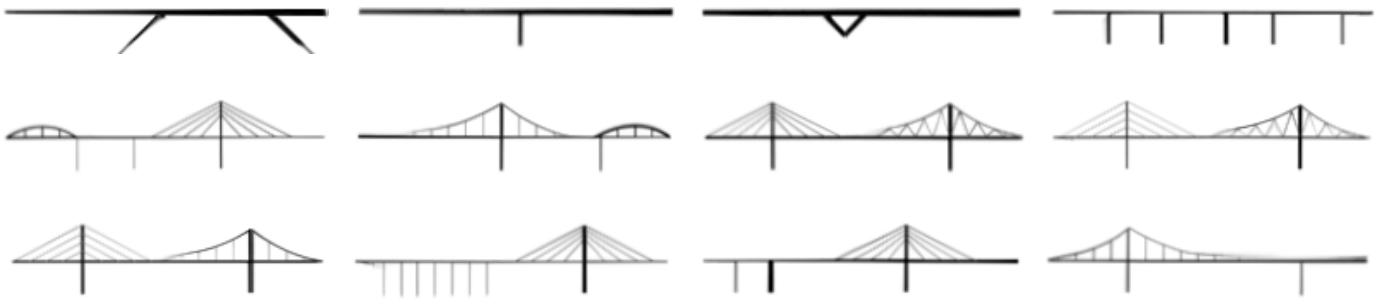

图 8 12 种技术可行的新桥型

Fig.8 Twelve new bridge types with feasible technology

这里新桥型是指数据集中未曾出现过，而神经网络根据算法创作出的，这代表了模型具有创新能力，比如上图第一张图“斜腿刚构桥”（现实中它是梁桥的子类）。某些桥型如独塔斜拉桥，现实中很常见。

而斜拉桥与悬索桥的纵向组合，据我所知未有工程师实践过，这是因为受力复杂或结构不尽合理。人类工程师在设计特大跨桥梁时，会尽可能让结构简单化。故这种组合桥型，不宜用于特大跨桥梁，但可用于城市的景观桥梁。

2.5 结果分析

数据集的桥型全是对称结构，而 PixelCNN 能够生成不对称的桥型，且并不是简单的叠合，而是将不同结构部件进行有机搭配，这点与生成对抗网络类似。

3 多模态模型是实现通用人工智能的路径

自回归模型的缺点是很明显的，它只是统计分析自身的历史数据，按照概率分布来确定下一个值，模型没有明白序列所表达的含义。如同仅通过音频来学习外语（无任何人指导、无任何语言场景），只能达到鹦鹉学舌的水平。

是样本空间之外的信息决定了空间中样本的意义，空间中样本是结果而不是原因。自回归模型仅对样本进行统计学习，治标不治本。故未来更先进的模型需要获取序列之外的信息才行。

人的智能同时具有回归和自回归模型的特性。生活中，人是根据现实场景来组织语言和行动的，这是回归模型行为。而 1945 年黄炎培关于“中国历史周期律”（是指中国历史上的政权经历兴衰治乱，往复循环呈现出的周期性现象）的著名思考，则是自回归模型行为。

多模态模型就是全面模仿人的智能，将回归和自回归模型结合，如同通过音频与场景之间的匹配来学习外语，从而掌握音频的含义，从而克服自回归模型的缺点。个人认为多模态模型应该是未来实现通用人工智能的路径。

4 结语

（1）PixelCNN 与生成对抗网络类似，比变分自编码器更有创造力，能够在人类原创桥型的基础上，将不同结构部件进行有机搭配，创造生成新桥型，一定程度上具有类似人类的原创能力，它能够打开想象空间，给予人类启发。

（2）PixelCNN 缺点是采样速度很慢，训练、实际部署运行对算力要求很高。

（3）自回归模型不能明白序列所表达的含义，而多模态模型将回归和自回归模型结合，从而能够理解序列。多模态模型应该是未来实现通用人工智能的路径。

参考文献

[1] 张洪俊，沈俊. BIM 技术前世今生和在常规桥梁设计中的应用[EB/OL]. 北京：中国科技论文在线 [2023-09-22]. http://www.paper.edu.cn/releasepaper/content/202309-51.

[2]张洪俊. 从变分自编码器隐空间中生成新桥型的尝试[EB/OL]. 北京：中国科技论文在线 [2023-11-06]. http://www.paper.edu.cn/releasepaper/content/202311-5.

[3]张洪俊. 从生成对抗网络隐空间中生成新桥型的尝试[EB/OL]. 北京：中国科技论文在线 [2023-12-25]. http://www.paper.edu.cn/releasepaper/content/202312-73.

[4] 弗朗索瓦·肖莱.Python 深度学习[M].北京:人民邮电出版社,2018.

[5] Aaron van den Oord, Nal Kalchbrenner, Koray Kavukcuoglu. Pixel Recurrent Neural Networks[J].arXiv preprint,2016,arXiv: 1601.06759.

[6] David Foster. Generative Deep Learning[M].2nd Edition. America: O’Reilly,2023.

[7] Aaron van den Oord, Nal Kalchbrenner, Oriol Vinyals, et al.Conditional Image Generation with PixelCNN Decoders[J].arXiv preprint,2016,arXiv: 1606.05328.

[8] Tim Salimans, Andrej Karpathy, Xi Chen, Diederik P. Kingma. PIXELCNN++: IMPROVING THE PIXELCNN WITH DISCRETIZED LOGISTIC MIXTURE LIKELIHOOD AND OTHER MODIFICATIONS[J].arXiv preprint,2017,arXiv: 1701.05517.

[9] Tfp.distributions.PixelCNN[EB/OL]. https://tensorflow.google.cn/probability/api_docs/python/tfp/distributions/PixelCNN, 2023-11-21.