DiffuseRoll：基于扩散模型的多轨、多类别音乐生成

1st 王宏飞中国传媒大学，北京，中国 f aywang@cuc.edu.cn

2nd 名 姓部门名称 所属机构名称 所属机构名称 城市，国家 电子邮件地址

3rd 名 姓部门名称 所属机构名称 所属机构名称 城市，国家 电子邮件地址

arXiv:2303.07794v1 [cs.SD] 14 Mar 2023

摘要——生成模型的最新进展在音乐生成方面取得了显著成果。然而，现有的大多数方法专注于生成单声部或主调音乐，而生成具有丰富属性的复调、多轨音乐仍是一项艰巨任务。本文提出一种利用扩散模型生成多轨、多属性交响音乐的新方法。具体而言，我们使用扩散模型生成钢琴卷帘表示，并将其映射为MIDI格式以供输出。为捕捉丰富的属性信息，我们引入一种颜色编码方案，将音符序列编码为颜色和位置信息，以表示音高、力度和乐器。该方案实现了离散音乐序列与连续图像之间的无缝映射。我们还提出一种后处理方法，以优化生成的乐谱，提升演奏效果。实验结果表明，与图中所示方法相比，在生成具有丰富属性信息的复调音乐方面，我们的方法优于最先进的方法。

索引词——交响乐生成、扩散模型、音乐表示、生成模型、钢琴卷帘

I. 引言

音乐生成是指利用计算机算法和技术创作新的音乐作品，这是人工智能领域的重要研究方向。音乐表示是指将音乐作品转换为计算机可处理的数字信号或数据结构，是音乐生成的基础和关键。音乐表示涵盖广泛，可包括音频信号的时频特征、乐谱的符号表示、音乐风格的元数据等。

由于音乐表示具有多样性，可用于预处理的方法和模型也各不相同。音乐音符可以被视为一种“语言”，即离散序列，因此可以采用基于Transformer的方法，例如 MuseFormer 和 SymphonyNet。

另一种方法是将乐谱转换为钢琴卷帘，并使用计算机视觉方法进行图像生成。此前的研究（如 MuseGAN）就采用了这种方法来生成多轨音乐，并将数据集中的乐器分为低音、鼓、吉他、钢琴和弦乐器。他们提出了一种生成对抗网络模型，用于生成钢琴卷帘，随后再经过后处理，生成 MIDI格式乐谱。

请在此处注明适用的资助机构。若无，请删除此内容。

尽管已有许多研究致力于推进音乐表示的发展，但仍面临诸多挑战。其中一个主要问题是如何将音乐特征与图像特征结合起来，以便在一张图中更好地描述乐器、力度等音乐的重要特征。因此，将图像处理中的一些方法和技术应用于音乐表示，可以提升音乐数据的表示能力和处理效果。

2021年，扩散模型在图像生成任务中表现出色，近期也已应用于文本和音频生成等其他领域。扩散模型在图示方法的交响乐生成方面展现出巨大潜力。然而，据我们所知，此前尚无研究使用扩散模型生成连续域多轨音乐。为填补这一空白，我们提出了一种使用扩散模型生成多轨音乐的新方法，有望生成高质量且多样化的多轨音乐。

我们的贡献如下：

- 我们提出了一种使用扩散模型生成多轨交响乐的新方法。我们的方法生成音乐的钢琴卷帘表示，随后将其转换为MIDI格式以供进一步处理。
- 我们提出了一种新的音乐表示方法，将MIDI格式转换为钢琴卷帘，并采用颜色编码方法对音高、力度和乐器等音乐特征进行编码，从而有助于在离散音乐序列与连续图像之间建立更好的映射。
- 我们提出了一种后处理方法，通过优化生成的乐谱来提升质量。通过大量实验，我们证明，与最先进的方法相比，我们的方法在生成音乐的多样性和逼真度方面表现更佳。总体而言，我们的工作推动了多轨音乐生成的发展，并为扩散模型在音乐生成领域的应用开辟了新可能。

II. 相关工作

A. 图形方法的音乐生成

在音乐生成领域，符号音乐生成被视为一种基础模型。Simon 等人[1] 于 2017年提出了 Performance-RNN，该模型利用长短期记忆（LSTM）循环神经网络 [2]生成古典钢琴音乐。另一方面，MusicVAE [3], 使用两层 LSTM 将音乐数据直接映射到潜在编码空间，然后通过单向循环神经网络对潜在空间进行解码来生成音乐。

生成对抗网络（GAN） [4],作为一种代表性的图形模型，自2016年起便被用于音乐生成。Mogren 等人 [5]提出了 C-RNN-GAN 模型，该模型通过循环神经网络处理连续音乐数据表示，并基于循环神经网络构建生成器和判别器模型，从而生成 MIDI格式的古典钢琴音乐。为捕捉音乐中的自重复，Harsh Jhamtani 和 Taylor Berg-Kirkpatrick 提出了自相似矩阵生成对抗网络（SSMGAN）[6]。引入自相似矩阵判别器大幅提高了生成器所生成音乐的真实性。2018年，Dong 等人 [7]推出了一种名为 MuseGAN 的生成对抗网络模型，用于生成编曲，该模型由三个子模型组成：即兴合奏、作曲和混合。通过是否共享噪声输入以及生成权重来控制各音轨之间的相关性。

B. 扩散模型

扩散模型在计算机视觉生成任务中表现出色，随后也在文本和音频等多个领域取得了令人瞩目的成果。去噪扩散概率模型（DDPM） [8] 是扩散模型领域的一项重要工作，在许多领域展现出巨大潜力。在音乐生成领域，谷歌此前提出了一种离散域扩散模型音乐生成算法，该算法采用自回归方式，通过反向过程学习生成潜在嵌入序列，并通过固定数量的迭代细化步骤实现并行生成。然而，目前尚无研究使用扩散模型在连续域中生成钢琴卷帘。与生成对抗网络模型相比，扩散模型在图像理解和可控性方面具有更多优势。因此，我们旨在研究将扩散模型用于钢琴卷帘生成。

III. 方法

A. 音乐表示编码器

音乐生成预处理：从 MIDI 转换为钢琴卷帘

在以往的音乐生成研究中，MuseGAN 提出了一种将 MIDI文件转换为钢琴卷帘的方法，并将音轨分为五类：低音、鼓、吉他、钢琴和弦乐器。该方法使用了 LMD数据集，这是一个流行音乐数据集。然而，在本研究中，我们使用一种交响乐数据集进行训练。由于其独特的管弦乐编配，该数据集比 MuseGAN 使用的流行音乐数据集更为复杂，也适合作为我们研究的参考。

受 MuseGAN 方法的启发，我们将 MIDI 文件中的 128 位乐器编码转换为五个音轨，并在同一张图像上用五种不同颜色表示。这样可以简化完整钢琴卷帘的生成过程，因为无需为每个音轨分别生成多个卷帘，从而避免增加计算复杂度。此外，音乐的颜色编码表示使我们能够在单个钢琴卷帘中用颜色表示乐器和音符强度。

钢琴卷帘生成的转换流程包括以下步骤：

1) Midi2array：我们采用刘家锋提出的方法，将音符序列表示为一个数组，其中每个音符都包含所有相关信息，包括音高、起始时间、时值、乐器、力度和速度。然后，我们将音符序列转换为相应的坐标点，记录音高、时间和音符长度。在此阶段，音符长度固定，每个音符在钢琴卷帘中表示为一个像素。如果一个音符是持续音高，我们就用多个音符来表示。生成的数组保存为 NPY文件。

Array2piano-roll 我们读取 NPY文件，得到一个二维数组，其中列表示每个音符像素的音高、乐器和强度特征，行表示音符总数。为了生成可变长度的音乐，我们按照像素位置对数组中的音符进行排列，如音符序列所示。我们将钢琴卷帘尺寸设为 128（对应音高）乘以自定义宽度 512；为方便训练，宽度设为 16 的倍数。表示一首乐曲所需的钢琴卷帘图像总数按 a * tpb / w 计算，其中 a 为较短音符的数量，tpb 为每拍刻数。

颜色编码音乐表示方法 颜色编码音乐表示方法将音乐的多种属性编码为 YUV色彩空间中的像素，其中 Y 表示亮度，对应 MIDI文件中的音符强度。由于 16 到 256 之间的 Y 值差异过大，我们将强度归一化到 50 至 100 的范围。该方法使我们能够将音符强度映射为钢琴卷帘的颜色，以表示音符的强弱。UV 分量表示色度，我们使用五种颜色表示不同的乐器类别。

我们还会在预处理过程中剔除力度值异常的音乐。例如，如果超过 80% 的音符都具有最大力度值，我们就认为该音乐存在问题并将其舍弃。这种方法提高了模型的鲁棒性以及生成结果的质量。最后，我们使用不同颜色绘制钢琴卷帘图像：绿色表示钢琴，蓝色表示打击乐，绿色表示木管乐器，青色表示弦乐器，品红色表示铜管乐器。

B. 建模

扩散模型是一种基于概率的生成模型，可用于生成与输入数据分布相似的新数据。其基本思想是通过迭代地向数据添加噪声，使数据逐渐接近高斯分布，从而降低数据处理的难度。在这个问题中，我们使用扩散模型生成钢琴卷帘。具体来说，我们将预处理后的钢琴卷帘作为输入，并生成新的钢琴卷帘。这里，我们将使用以下框架。

1) 前向过程：首先，我们需要定义条件概率分布 $q(\mathbf{x}_t|\mathbf{x}_0)$，其中 $\mathbf{x}_0$ 表示输入的钢琴卷帘图像。该条件分布可通过以下公式定义。

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \tag{1}$$

$$q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1}) \tag{2}$$

其中，$\mathbf{x}_{t-1}$ 是上一个时间步的输出，$\{\beta_t \in (0,1)\}_{t=1}^{T}$ 是用于控制时间步的方差调度。最后，$\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma})$ 表示均值为 $\boldsymbol{\mu}$、协方差为 I 的高斯分布。

由上式推导可得：

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{3}$$

$$\alpha_t = 1 - \beta_t \tag{4}$$

$$\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i \tag{5}$$

前向过程使用公式 3。

2) 反向过程：

$$\tilde{\boldsymbol{\mu}}_t = \frac{1}{\sqrt{\alpha_t}}\Big(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_t\Big) \tag{6}$$

其中 $\tilde{\boldsymbol{\mu}}_t$ 是要添加到各步骤中的噪声均值，$\boldsymbol{\epsilon}_t \sim \mathcal{N}(0, \mathrm{I})$

我们希望.....我们通过最小化交叉熵来训练模型，并将其作为学习目标。

$$\begin{aligned} L_{\mathrm{CE}} &= -\mathbb{E}_{q(\mathbf{x}_0)} \log p_\theta(\mathbf{x}_0) \\ &\le -\mathbb{E}_{q(\mathbf{x}_{0:T})} \log \frac{p_\theta(\mathbf{x}_{0:T})}{q(\mathbf{x}_{1:T}|\mathbf{x}_0)} \\ &= \mathbb{E}_{q(\mathbf{x}_{0:T})}\Big[\log \frac{q(\mathbf{x}_{1:T}|\mathbf{x}_0)}{p_\theta(\mathbf{x}_{0:T})}\Big] = L_{\mathrm{VLB}} \end{aligned} \tag{7}$$

此处，T是时间步的数量，$|\boldsymbol{x}|_2$ 表示 $\boldsymbol{x}$的 $L2$参数化。我们可以使用随机梯度下降来最小化损失函数，并根据需要

$$\begin{aligned} L_{\mathrm{VLB}} = \mathbb{E}_q[&\underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_T|\mathbf{x}_0) \parallel p_\theta(\mathbf{x}_T))}_{L_T} \\ &+ \sum_{t=2}^{T} \underbrace{D_{\mathrm{KL}}(q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0) \parallel p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t))}_{L_{t-1}} \\ &\underbrace{-\log p_\theta(\mathbf{x}_0|\mathbf{x}_1)}_{L_0}] \end{aligned} \tag{8}$$

C. 后处理

在后处理中，我们将生成的图像转换为钢琴卷帘。首先，将图像转换为音符序列，图像上的每个点都代表一个音符。在这一步中，我们会滤除一些过于“密集”的音符。由于这是多轨交响乐，一个片段中的音符过多会导致音符之间不和谐，因此删除其中一些音符，以确保片段的和谐。然后，我们将音符序列转换为 MIDI 并输出。

IV. 实验

我们通过主观和客观评估来衡量所提出模型在钢琴卷帘生成方面的性能。我们使用包含古典交响乐作品的 MIDI 文件交响乐数据集训练模型，并使用多种指标评估生成的钢琴卷帘。

A. 客观评估

我们使用了多种客观指标评估生成的钢琴卷帘，包括以下几项：

准确率：我们计算了生成的钢琴卷帘中与对应真实钢琴卷帘中的音符相匹配的比例。

精确率和召回率：我们计算了生成的钢琴卷帘相对于真实钢琴卷帘的精确率和召回率。

F 值：我们计算了生成的钢琴卷帘相对于真实钢琴卷帘的 F 值。

信息熵：我们计算了生成的钢琴卷帘的信息熵，以评估生成音符的多样性。

Inception 分数：我们计算了生成的钢琴卷帘的 Inception 分数，以评估其多样性和质量。

客观评估结果表明，在准确率、精确率、召回率和 F 值方面，我们提出的模型优于 MuseGAN。生成的钢琴卷帘具有较高的信息熵，表明生成的音符多样性较高。生成的钢琴卷帘的 Inception 分数也较高，表明其质量较高且具有多样性。

B. 主观实验

我们进行了听力测试，以评估生成的钢琴卷帘的质量。我们请 20 位经验丰富的音乐家对生成的钢琴卷帘进行评分，评分范围为 1（较差）到 5（优秀）。我们随机

选取了 10 个生成的钢琴卷帘和 10 个真实的钢琴卷帘，并按随机顺序播放给参与者听。我们请参与者对钢琴卷帘的整体质量、音乐性和真实性进行评分。

C. 结论

我们提出的基于扩散建模的钢琴卷帘生成模型，在主观和客观评估中均取得了高质量的结果。这些结果证明了我们提出的方法在生成高质量钢琴卷帘方面的有效性。未来的工作将着重于扩展数据集，并纳入更复杂的建模技术，以进一步提升所提模型的性能。

参考文献

[1] Ian Simon 和 Sageev Oore，《Performance RNN：生成具有表现性时序和力度变化的音乐》，2017年。[2] Sepp Hochreiter 和 J¨urgen Schmidhuber，《长短期记忆》，Neural Computation，第 9 卷，第 8 期，1735–1780 页，1997 年。[3] Adam Roberts、Jesse Engel、Colin Raffel、Curtis Hawthorne 和 Douglas Eck，《用于学习音乐长期结构的层次潜向量模型》，载于《国际机器学习会议》。PMLR，2018 年，4364–4373 页。[4] Ian Goodfellow 、 Jean Pouget-Abadie 、 Mehdi Mirza 、 Bing Xu 、 David Warde-Farley、Sherjil Ozair、Aaron Courville 和 Yoshua Bengio，《生成对抗网络》，Communications of the ACM，第 63 卷，第 11 期，139–144 页，2020 年。[5] Olof Mogren，《C-RNN-GAN：采用对抗训练的连续循环神经网络》，arXiv 预印本 arXiv:1611.09904，2016 年。[6] Harsh Jhamtani 和 Taylor Berg-Kirkpatrick，《使用生成对抗网络对音乐生成中的自重复进行建模》，载于音乐发现机器学习研讨会，ICML，2019年。[7] Hao-Wen Dong、Wen-Yi Hsiao、Li-Chia Yang 和 Yi-Hsuan Yang，《MuseGAN：用于符号音乐生成和伴奏的多轨序列生成对抗网络》，载于 AAAI人工智能会议论文集，2018 年，第 32 卷。[8] Jonathan Ho、Ajay Jain 和 Pieter Abbeel，《去噪扩散概率模型》，神经信息处理系统进展，第 33 卷，6840–6851 页，2020 年。

此图“fig1.png”提供以下格式的文件：“PNG”：

http://arxiv.org/ps/2303.07794v1