arXiv:2301.02385v1 [cs.SD] 6 Jan 2023

多流派 音乐 Transformer——创作完整篇幅的音乐作品 e

Abhinav Kaushal Keshari（普渡大学）

摘要

在音乐生成任务中，艺术性发挥着重要作用，也是人工智能面临的一大挑战。此前，研究者通过对抗训练（Dong 等人，2018）创作新的音乐作品，并对音乐中的多样性（节拍、速度、音乐分轨）进行兼容性建模（Huang等人，2021），展示了学习这项任务的出色范例。不过，这些工作仅限于生成混搭作品，或从速度和调的分布中学习特征，以生成相似的模式。复合词 Transformer（Hsiao 等人，2021）能够将音乐生成任务表示为序列生成问题，其中音乐事件由复合词定义。这些音乐事件能够更准确地描述音符走向、和弦变化、和声以及艺术性。本项目旨在实现一种多风格变换器，通过更具自适应性的学习流程来学习创作音乐作品，并应对更具挑战性的任务，其中还会考虑作品的流派或曲式。我们构建了一个多流派复合词数据集，并实现了线性 Transformer（Katharopoulos等人，2020），并在该数据集上进行训练。我们将其称为多风格变换器。它能够生成完整篇幅的新音乐作品，其多样性可与原始曲目相媲美。该模型的训练速度比文中讨论的其他模型快 2 至 5 倍。

1. 相关工作

尽管人工智能在自然语言生成（NLG）任务中取得了巨大成功，但艺术这一因素仍使生成结果与人类的表现有所不同。就自然语言生成而言，这类似于计算机生成的文章与小说、传记等艺术作品之间的差别。音乐始终离不开艺术因素；尽管我们能够通过对抗网络创作音乐作品，或使用监督学习混合分轨，但这些方法得到的结果仍与原创音乐作品大不相同，下面将对此展开讨论。

1.1. 使用生成对抗网络进行音乐生成

生成对抗网络（GAN）在文本、视频和图像生成方面取得了显著进展。人们也作出了类似尝试，将神经网络应用于音乐艺术领域。MuseGAN(Dong etal., 2018)提出了一种生成多轨音乐的新颖模型。

截至 2018 年，利用人工智能创作音乐的进展已能够生成

- 单轨（单声部）音乐
- 通过按时间顺序组合多个单声部旋律生成的多轨（复调）音乐

音乐通常是一种由多种乐器合奏的艺术，因此需要采用多轨形式；而且，由于音乐音符由和弦、琶音或旋律构成，按时间顺序排列的设定会使其难以泛化。

该论文（Dong 等人，2018）通过讨论神经网络模型在生成真实音乐方面的当前技术缺陷，以及这些缺陷与现实世界音乐的关系，探讨了这一挑战。

1. 音乐是一种时间艺术，具有连贯性、节奏、张力和情感流动等特点。因此，它需要一个**时间模型**。

2. 音乐作品通常涉及不同乐器之间的互动，使作品具有和声性。要解决这一问题，需要一个**作曲模型**。

3. 音符由和弦、琶音或旋律构成，并随时间展开；因此，按时间顺序生成音符并不合适。为解决这一问题，论文提出以小节（时间片段）而非音符作为创作的基本单位，并使用转置卷积神经网络逐小节生成音乐，以学习平移不变模式。

这篇论文（Dong 等人，2018）在人工创作逼真音乐的能力，以及采用结合时间与作曲模型的生成对抗框架方面做出了贡献。简而言之，其贡献如下：

- 首个基于生成对抗网络的多轨序列生成模型。
- 首个能够生成多轨复调音乐的模型。
- 同一模型也可用于音乐伴奏。
- 为多轨钢琴卷帘谱创建了一个新的 Lakh 钢琴卷帘谱数据集（LPD）
- 针对人工音乐领域的未来研究，提出了一套新的客观指标。

所提出的 MuseGAN 模型包含两个子网络生成器：Gtemp（时间结构生成器）和 Gbar（小节生成器），共同构成整体生成器：

$$G(z) = \left\{G_{bar}(G_{temp}(z)^{(t)})\right\}_{t=1}^{T}$$

其中，z 是输入噪声向量。该模型的优势在于能够生成具有类似和弦的音程（从时间模型中学习特征）以及吉他、钢琴和弦乐之间存在音高重叠的旋律（从作曲模型中学习特征）。

该模型通过提出三种不同的生成器模型（即兴合奏、作曲家和混合型），对声部间依赖关系进行建模，从而引入了多轨机制；但作者是基于对流行音乐创作的理解提出这些模型的。这可能会限制生成器对更广泛音乐类型的探索，使其难以实现通用化。还值得一提的是，这项研究依赖多轨间依赖关系，却没有探讨这些声部之间的兼容性，而兼容性能够显著提升生成音乐的质量。下一篇论文将讨论这一问题。

1.2. 对分轨兼容性建模以生成音乐混搭（Huang等人，2021）

音源分离（Jansson等人，2017年；D´efossez等人，2019年）使得利用分离出的人声、鼓、钢琴等分轨来生成音乐混搭成为可能。挑战在于如何创作出各分轨彼此兼容的音乐。本文构建了一个混搭生成流程，并训练模型，通过自动学习调整调和速度（现实中优质混搭作品的特征）来预测兼容性。

针对和声兼容性训练的通用模型（Bernardes等人，2017年；Macas等人，2018）未能考虑细微特征，也无法处理来自不同样本的意外混音，而这在这一艺术领域相当常见。另一个问题是，像 Neural Loop Combiner 这样的音频兼容性模型（Chen等人，2020）缺少人声音源，且音乐流派多样性不足。

作者通过重新组合音源分离前原有的分轨组合，设计了一种**自监督学习**模型，并将其用作真实标注示例。为避免模型过度极化，**半监督学习**应运而生，其中包括通过混合不同分轨生成若干随机混搭作品，并将其视为未标记实例。针对离群值的标签平滑正则化（Zheng 等人，2017年）用于在计算损失时，为未标记数据赋予均匀分布。这有助于正则化。

最终架构由 3 个模块组成：

1.**音乐源分离：**使用 MSS算法（Jansson等人，2017年）提取不同的分轨，包括人声、鼓、贝斯和其他音轨。

2. **混搭数据库（MashupDB）：**使用 Madmom（B¨ock等人，2016年）从音乐片段中提取调、速度和强拍等特征信息。利用这些特征和分离后的音轨组合，可创建一个混搭数据库，作为混搭生成过程中和声音轨或打击乐音轨的候选素材。

3. **混搭生成：**该模块使用 MashupDB 中的候选音轨，并调整调和速度，在三种条件下生成混搭作品——原始、匹配和不匹配。

该模型（Huang等人，2021）定义为 $p(y|V,\ H,\ P)$ 其中 V，H和 P分别是对应分轨的输入信号：人声、和声和打击乐。输出概率 p 用作混搭兼容性，$y \in \{0, 1\}$ 表示好或坏。

该模型（Huang等人，2021）的实现尝试模拟学习兼容性的过程，以创作新的混搭作品，并通过在多个不同数据集之间进行交叉验证，提供客观和主观评估。这种方法更容易实现，因为模型能够提取不同的分轨和特征，并构建自己的混搭候选作品。这也使模型的训练过程不依赖人工标注数据。由于在监督学习中同时加入了负样本数据和正样本数据，该模型也具有较强的鲁棒性。其音乐覆盖范围也很广，而且音源分离步骤使模型更容易扩展到不同音乐流派进行训练。

但当前模型设计在生成混搭作品时，未能有效嵌入不同的分轨，并且依赖对调和速度进行调整。目前的实现为确保兼容性，采用了固定范围的调和速度差异，却没有详细说明这些数值是如何得出的。尽管设定范围可以避免大幅移调和时间拉伸，但模型结果对正向标注数据（原始数据）的排序高于未标记数据，这可能会引发

对灵活性的担忧。该模型面临的另一项重大挑战是训练时间过长：使用 NVIDIA Tesla-V100 GPU 约需 3 天，而采用 Transformer 模型则能显著缩短训练时间。

1.3. 音乐 Transformer

借助最先进的神经网络，我们通过定义关于速度、节拍或兼容性的匹配规则，成功学习了音乐中的特征。在上一篇论文中，我们还尝试借助监督学习来学习兼容性。不过，该模型存在偏差：它偏向于匹配调或速度的情况，而且缺乏泛化能力。复合词 Transformer (Hsiao 等人，2021) 将音乐视为事件序列，并使用 Transformer（一种神经序列模型）(Vaswani 等人，2017年) 来生成新的音乐序列。

音符可以通过其音高、和弦、小节、时值、力度（力度变化）和位置（起音时间）来描述。如果将这些视为词元，我们就可以把音乐定义为词元序列，而这些词元属于预先定义的词汇表。由于音乐包含多个方面，特定类型的词元只能捕捉某些特征，例如旋律、节奏和和声。迄今为止，所有神经网络都将这些词元视为同等的，因此缺乏异质性。

复合词 Transformer（Hsiao 等人，2021）以一种概念上不同的方式生成音乐，因为它允许词元具有特定类型，并拥有各自的属性。词元可以是音符类型（音高、时值），也可以是节拍类型（新拍子的开始、小节）。随后，我们将此类词元组合起来定义音乐事件，从而捕捉词元之间的共现关系。这种词元组合称为复合词。因此，现在我们可以将一段音乐作品 (X) 表示为复合词序列(S)，(cp) 或 $S = g(X) = \{cp_t\}_{t=1}^{T}$ 其中 $g(.)$ 是将音乐转换为按时间排序的音乐事件序列的转换函数，而 T 是音乐序列的长度。

理论上，该模型在离散时间动态有向超图上进行学习。考虑一个图 $G = (V, E)$（图 1），其中顶点（V）是词元，边（E）是词元序列。顶点集合可定义为一个复合词，而该图中的超边表示复合词序列。图 1 $v1$, $v2$, $v5$ 是词元，边 $E1$ 定义了一个词元序列，而 $e1$, $e2$定义了一条超边（连接两个以上的节点）。从一条超边转移到另一条超边，便定义了我们试图学习的作品词序列。

我们尝试使用 Transformer 来学习下一个音乐事件或复合词（词元的组合）。Transformer 中的自注意力部分会学习音乐序列中各元素之间的依赖关系，而不同的前馈

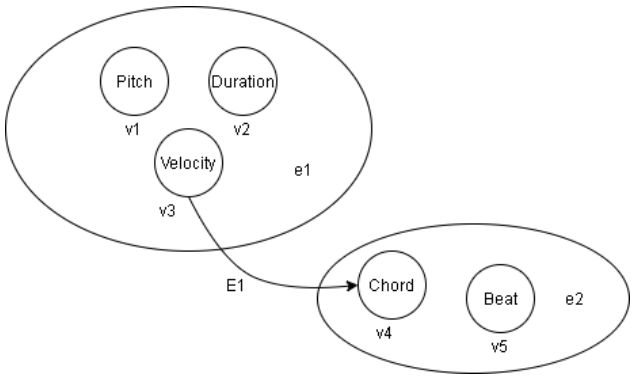


图 *1.* 音乐空间的图形表示

头则用于不同类型的词元。简而言之，该实现将词元组合成复合词，然后对这一复合词序列进行序列建模，其主要贡献如下：

- 创作一首完整歌曲长度的流行钢琴音乐。
- 复合词序列建模采用线性 Transformer，在质量方面取得了最先进的结果，训练和推理速度快 5–10 倍。
- 音乐被定义为动态有向超图。

该模型的核心是在每个时间步生成一个新的音乐事件，或生成一组可组合成复合词的词元，但这依赖于一个假设：任意两个音乐事件都不能同时发生。Transformer解码器生成的新超边会在检测到特定词元类型的事件后，将其他词元标记为[ignore] 。这会限制音乐生成任务吗？此外，该模型仅使用流行音乐进行训练，这限制了 Transformer 的表达能力。

2. 实现

复合词 Transformer(Hsiao 等人，2021) 能够将音乐生成任务表示为序列生成挑战，其中涉及由复合词定义的音乐事件。基于这种表示，我们实现了一个神经模型，通过更具自适应性的学习过程来学习创作音乐作品；这一过程涉及更具挑战性的任务，还会考虑作品的流派或曲式。这使注意力驱动的序列学习过程能够融入音乐艺术的丰富性。我们将这个模型称为 多流派音乐 *Transformer* ，以下是实现步骤：

- **构建数据集**：这包括为不同音乐流派的歌曲生成复合词词典。

- **实现 Transformer 模型**：我们实现 Transformer 类、训练步骤以及用于推理的生成逻辑。
- **自适应学习**：我们通过使用规模更小、涵盖多种流派的数据集进行训练，使调优后的模型能够适应不同情况。

2.1. 构建数据集

为了让 Transformer 能够进行更加泛化的学习，需要使用包含不同流派/风格音乐作品的钢琴卷帘数据集对其进行训练。该数据集应基于复合词（Hsiao 等人，2021），将不同的音乐词元表示为用于序列建模的组合单元，这与传统音乐数据集（MIDI、REMI）有所不同。

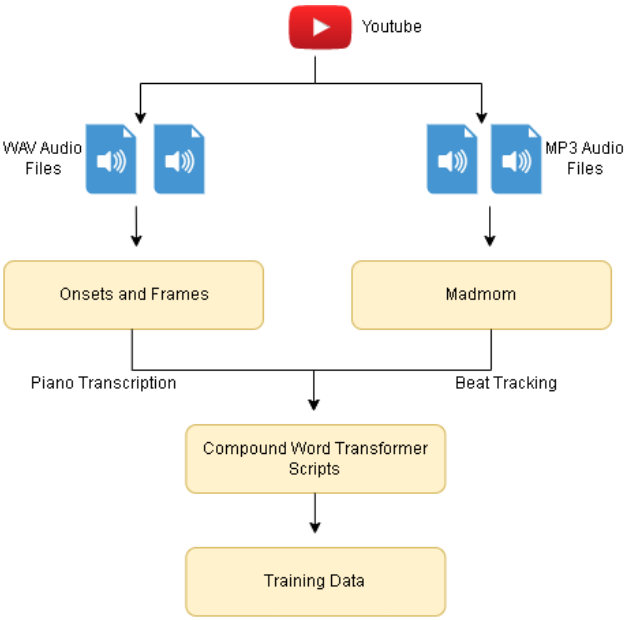


图 2.数据集构建流水线

为此，我们需要构建一个数据集：选取音乐片段，并使用 Onsets and Frames 将其转换为钢琴卷帘（Hawthorne等人，2017年）。然后，使用音乐信号处理库 Madmom 从这些歌曲中提取弱拍和节拍信息（B¨ock 等人，2016年）。最后，使用复合词 Transformer代码库提供的数据集生成脚本，将这些元数据转换为复合词表示[1]。此外，这也扩充了 AILabs.tw Pop1K7数据集（Hsiao 等人，2021），该数据集目前仅包含流行音乐。图 2 展示了创建新数据集的流水线。

按照上述流程，我们成功创建了一个复合词（Hsiao 等人，2021）数据集，其中包含来自 3 种不同音乐流派的 150 首音乐作品的钢琴卷帘，包括电子舞曲（EDM）、独立音乐和嘻哈音乐。

[1]https://github.com/YatingMusic/compound-word-transformer/blob/main/dataset/Dataset.md

2.2. 实现 Transformer 模型

我们实现了一个线性 Transformer（Katharopoulos等人，2020），以解决长序列依赖问题。这是音乐生成中的一个重要因素，因为整首音乐作品中存在上下文或节奏。在 Transformer解码器中设置独立的前馈头，有助于降低独立词元的损失。通过添加额外的词元类型，模型可以扩展到音乐的其他视角（例如流派、曲式或特定的和弦进行）。我们以通用方式实现了 Transformer 模型，用户可以自行定义词元采样模型和词元嵌入模型，并且这些模型可以扩展到任意数量的词元类型。图 6 展示了各前馈头观测到的损失。这表明，只需为模型添加一个独立的前馈头，就能让模型学习新的词元（用于流派/风格/曲式）。

2.2.1. 词元嵌入

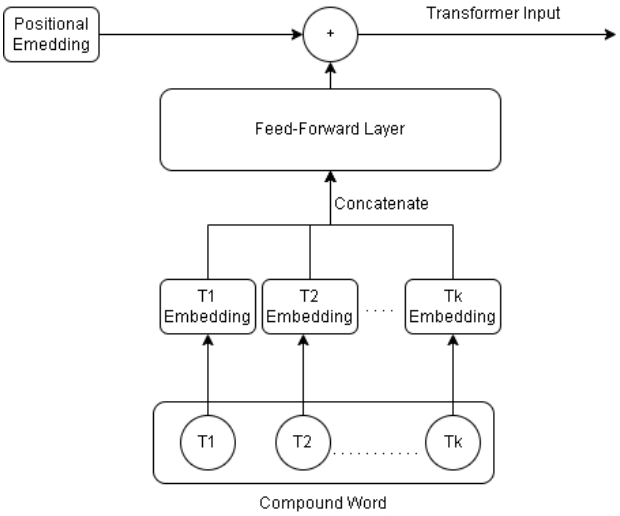


图3。展示了每个词元在与位置编码结合之前如何独立嵌入。这里T_1，T_2...T_k 我们的 Transformer 有 K 种不同的词元，每种词元都有自己的嵌入函数和维度。我们假设 Transformer 支持 K 种类型的词元。

Transformer 的输入需要在输入序列元素的嵌入向量上添加位置编码。由于序列中的每个元素都是由不同词元组合而成的复合词（Hsiao 等人，2021），我们会分别嵌入每个词元（从而支持自适应大小），然后将它们拼接起来。采用自适

应词元大小后，词汇表较小的词元类型便可使用较小的嵌入维度；将所有词元拼接后，我们的模型得到的嵌入维度为 512。词元嵌入的详细步骤请参见图3。

2.2.2. 词元采样

在推理过程中，采样对于避免退化和提升多样性至关重要。为避免退化，我们采用核采样（Holtzman 等人，2019年），这是一种随机温度控制过程。该方法从累积概率质量超过阈值的最小词元子集中进行采样。我们还为每个词元设置独立的采样策略，分别定义不同的阈值 p 和不同的温度参数τ （Ackley 等人，1985年），在采样前重新调整概率。我们沿用了复合词 Transformer（Hsiao 等人，2021）的推理实现，并对其进行了调整 τ，使和弦对应的数值更高，从而产生更多样的和弦进行。图 4 展示了采样过程以及 Transformer 中每个词元对应的独立前馈层。

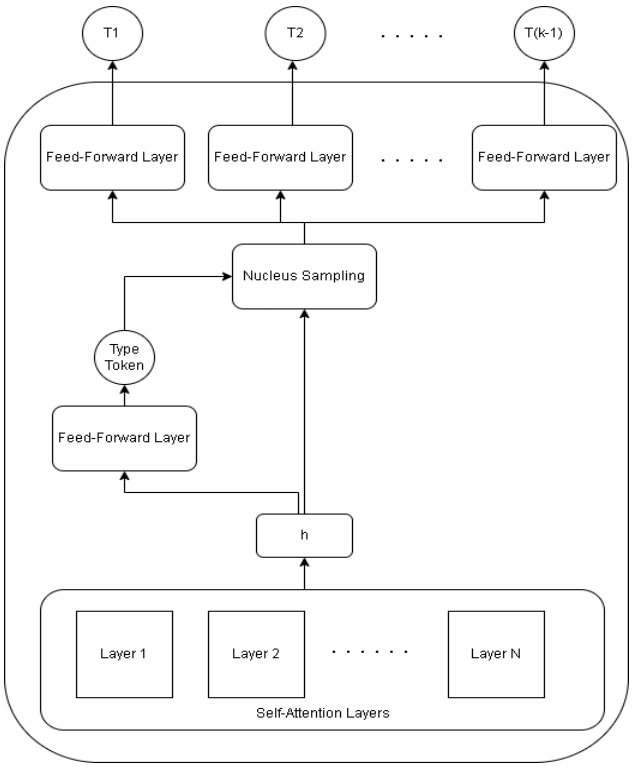


图 *4.* 具有 N 个自注意力层、并为每种词元配置独立前馈头的 Transformer。我们首先预测特定时间步的词元类型，然后进行核采样，再预测其余词元。

2.3. 自适应学习

定义模型后，接下来的重要步骤是实现训练步骤。为了支持通用 Transformer 中可扩展的词元定义，我们将训练步骤设计为模块化且通用，可适应数量可变的词元类型。这样便于添加新词元，并能独立监控相应损失的梯度下降优化。

我们针对两种不同条件并行训练模型。第一组训练使用原始 AIL-abs.tw Pop1K7 数据集（Hsiao 等人，2021）。第二组训练旨在为 Transformer 提供多流派学习环境，训练所用词典由三种不同流派（EDM、独立音乐、嘻哈）生成。

3. 评估与结果

训练多流派 Transformer 的主要目标，是为其提供一个比原始纯流行音乐数据集更加丰富多样的数据集。借助数据集构建流水线，我们成功创建了一个方差更高的词元集，使模型具备更广泛的表达能力。图 5 展示了所用两个数据集的词元对比。

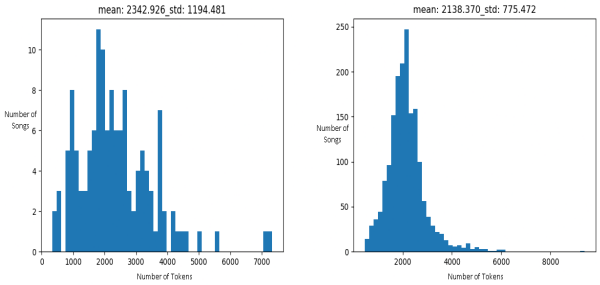


图 *5*。左图展示生成的多流派数据集中歌曲的词元分布，右图展示 AILabs.tw Pop1K7数据集中的类似分布（Hsiao 等人，2021）。

在分别使用两个数据集训练模型后，我们还观察到（参见图 6），单个词元损失与总平均损失相似，表明模型正在收敛。此外，使用多流派数据集时，梯度下降过程更加平缓，呈现出更稳定的变化趋势。

我们使用 12 层自注意力层、8 个前馈头、512 的模型维度和 4 的批量大小训练模型，共训练 180 轮，耗时约 17 小时。随后，我们使用训练好的模型生成了 20 首新的完整音乐作品，平均每首歌曲的推理时间为 12.56 秒，比复合词 Transformer 更快，但每首歌曲的平均词元数量略少。表1展示了更详细的比较。

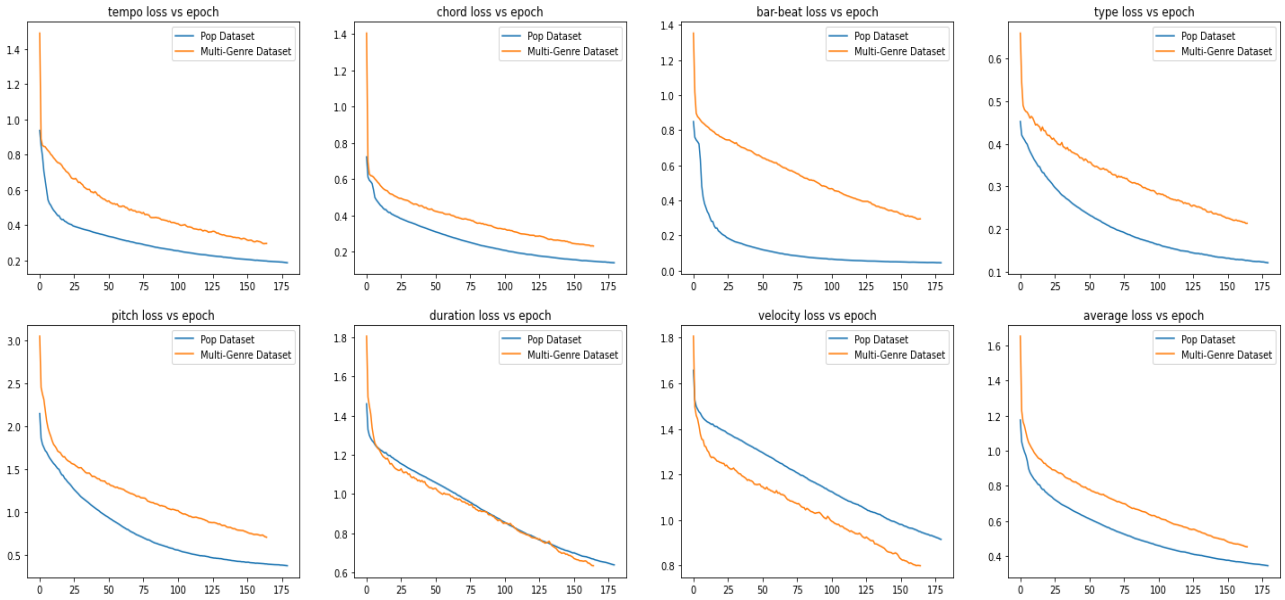
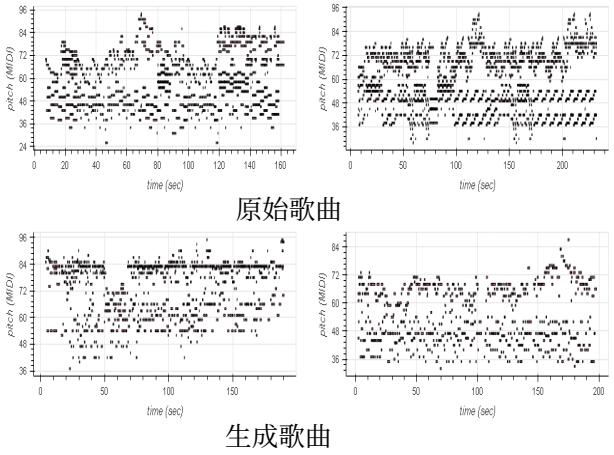


图 *6*。不同词元类型的损失与训练轮次关系图。最后一幅图对应所有不同词元类型的平均损失。

为了对生成的音乐作品进行定性评估，我们比较了（图 7）这些作品的钢琴卷帘与用于训练模型的原始曲目的钢琴卷帘。



原始歌曲

生成歌曲

图 *7*。原始歌曲和生成歌曲的钢琴卷帘。可以看出，生成歌曲的内容丰富且完整，与一些原始曲目相似。

4. 结论

在本项目中，我们使用训练好的 Transformer 将音乐生成为一系列音乐事件。我们借鉴复合词的定义（Hsiao 等人，2021年），通过组合多个词元来定义音乐事件。这种组合大幅缩短了序列长度，并提升了长距离学习能力，也显著缩短了模型的训练和推理时间。我们还利用每个词元都有独立前馈头用于预测这一特性，使模型能够扩展以支持词典中新增的词元类型。这样便能轻松地为该 Transformer 添加任何新词元，用于表示曲式、和弦进行等。此外，我们创建了一个全新的数据集，其中包含多流派复合词词典，并据此训练模型，为其提供更具自适应性的学习环境。生成的作品包含丰富的音乐事件，质量也很高。

表1。多风格变换器和复合词 Transformer 的定量评估结果。复合词 Transformer 的结果来自论文中的实现（Hsiao 等人，2021）。

模型	训练时间	GPU	推理时间（/首歌曲）	平均词元数（/首歌曲）
多风格变换器	17 小时	9.8GB	12.56 秒	9190
复合变换器	1.3 天	9.5GB	19.8 秒	9546

参考文献

Ackley, D. H.、Hinton, G. E. 和 Sejnowski, T. J. 玻尔兹曼机的学习算法。《认知科学》，9(1):147–169，1985年。URL https: // www.sciencedirect.com/science/ article/abs/pii/S0364021385800124。Bernardes, G.、Davies, M. E. 和 Guedes, C. 一种分层谐波混音方法。载于国际计算机音乐多学科研究研讨会，第151–170页。施普林格，2017年。URL https://link.springer.com/chapter/ 10.1007/978-3-030-01692-0_11。B¨ock, S.、Korzeniowski, F.、Schl¨uter, J.、Krebs, F. 和 Widmer, G. Madmom：一个新的 Python 音频与音乐信号处理库。载于第24届 ACM 多媒体国际会议论文集，第1174–1178页，2016年。URL https://dl.acm.org/doi/abs/10. 1145/2964284.2973795。Chen, B.-Y.、Smith, J. B. 和 Yang, Y.-H. 神经循环片段组合器：用于评估循环片段兼容性的神经网络模型。arXiv 预印本 arXiv:2008.02011，2020。URL https://arxiv.org/abs/2008.02011。D´efossez, A.、Usunier, N.、Bottou, L. 和 Bach, F. 波形域中的音乐音源分离。arXiv 预印本 arXiv:1911.13254，2019。URL https:// arxiv.org/abs/1911.13254。Dong, H.-W.、Hsiao, W.-Y.、Yang, L.-C. 和 Yang, Y.-H. MuseGAN：用于符号音乐生成与伴奏的多轨序列生成对抗网络。载于第三十二届 AAAI 人工智能会议，2018年。URL https://www.aaai.org/ocs/index.php/ AAAI/AAAI18/paper/viewPaper/17286。Hawthorne, C.、Elsen, E.、Song, J.、Roberts, A.、Simon, I.、Raffel, C.、Engel, J.、Oore, S. 和 Eck, D. Onsets and Frames：双目标钢琴转录。arXiv 预印本 arXiv:1710.11153，2017年。URL https:// arxiv.org/abs/1710.11153。Holtzman, A.、Buys, J.、Du, L.、Forbes, M. 和 Choi, Y. 神经文本退化的奇特案例。载于国际学习表征会议，2019年。URL https://openreview.net/forum? id=rygGQyrFvH。Hsiao, W.-Y.、Liu, J.-Y.、Yeh, Y.-C. 和 Yang, Y.- H. 复合词 Transformer：学习在动态有向超图上创作整首歌曲音乐。载于 AAAI 人工智能会议论文集，第35卷，第178–186页，2021年。网址 https://ojs.aaai.org/index.php/AAAI/article/view/16091。

Huang, J., Wang, J.-C., Smith, J. B., Song, X., and Wang, Y. 对分轨的兼容性进行建模，以生成音乐混搭作品。载于*AAAI* 人工智能会议论文集，第35卷，第187–195页，2021年。网址 https://ojs.aaai.org/index.php/AAAI/article/view/16092。

Jansson, A., Humphrey, E., Montecchio, N., Bittner, R., Kumar, A., and Weyde, T. 使用深度U-Net卷积网络进行歌声分离。载于第*18*届国际音乐信息检索学会会议，第23–27页，2017年。网址 https://openaccess.city.ac.uk/id/eprint/19289/1/7bb8d1600fba70dd79408775cd0c37a4ff62.pdf。

Katharopoulos, A., Vyas, A., Pappas, N., 和 Fleuret, F. Transformer 就是循环神经网络：采用线性注意力的快速自回归 Transformer。在国际机器学习会议，第 5156–5165 页。机器学习研究论文集，2020年。网址：http://proceedings.mlr.press/v119/katharopoulos20a.html。

Macas, C., Rodrigues, A., Bernardes, G., 和 Machado, P. Mixmash：用于音乐混搭创作的可视化系统。在*2018*年 第*22*届信息可视化国际会议（*IV*），第 471–477 页。IEEE计算机学会，2018。网址：https://www.computer.org/csdl/proceedings-article/iv/2018/720200a471/17D45XvMcd9。

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., 和 Polo-sukhin, I. 注意力就是你所需的一切。在神经信息处理系统进展，第5998–6008页，2017年。网址：https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper pdf.

Zheng, Z.、Zheng, L. 和 Yang, Y.。生成对抗网络生成的无标签样本可在体外改善行人重识别基线。在*IEEE* 国际计算机视觉会议论文集，第 3754–3762 页，2017年。网址：https://openaccess.thecvf.com/content_iccv_2017/html/Zheng_无标签_样本_生成_ICCV_2017_论文.html。