基于深度注意力对齐网络的不完整歌词旋律生成

Gurunath Reddy M*†‡、Zhe Zhang*†、Yi Yu*、Florian Harscoet*‡、Simon Canales*‡ 和 Suhua Tang§
*日本国立信息学研究所数字内容与媒体科学研究部、综合研究大学院大学 yiyu
@nii.ac.jp§电气通信大学信息学与工程研究生院计算机与网络工程系

arXiv:2301.10015v1 [cs.SD] 23 Jan 2023

摘要—**我们提出了一种深度注意力对齐网络，旨在以不完整歌词为输入，像人类创作音乐那样自动预测歌词和旋律。最重要的是，我们采用编码器-解码器方式训练深度神经歌词到旋律网络，使其在给定不完整歌词（少量关键词）时预测可能的歌词-旋律组合。在歌词到旋律生成过程中，利用注意力机制将预测的歌词与旋律对齐。定性和定量评估指标表明，所提出的方法确实能够根据一段不完整的种子歌词，生成恰当的歌词及相应旋律，用于创作新歌曲。**

I. 引言

自动音乐生成旨在借助机器辅助环境创作有意义的音乐，这一直是人工智能和音乐领域的一项具有挑战性的研究课题。过去几年，我们见证了自动音乐生成领域的研究工作日益增多。随着互联网上海量音乐资源的涌现，隐马尔可夫模型 [1],图模型 [2], 和深度学习模型 [3]–[7], 等数据驱动方法展现出音乐创作的潜力。关于无条件音乐生成的研究已相当丰富，而在给定完整或不完整歌词的情况下根据歌词生成旋律的研究则较少。条件旋律生成研究明显较少，主要原因可归结为：i）缺乏歌词-旋律配对数据集；ii）一段歌词创作可能对应多种旋律表示，因此难以学习歌词与旋律之间的关联；iii）生成的旋律难以进行客观评估。

本研究聚焦于音乐生成中最具挑战性的方面：使深度学习模型能够在输入少量词语后，发现恰当的歌词和富有意义的旋律。据我们所知，这是首次尝试根据不完整歌词生成旋律，适用于没有音乐知识但希望创作歌曲的人。具体而言，我们提出了歌词转向量（lyric2vec）模型，该模型基于大型英语流行歌曲数据集进行训练，以获取歌词在音节、词和句子层级上的稠密表示。我们提出的歌词转旋律网络（LTMN）采用编码器-解码器架构，并基于深度注意力机制对齐进行训练。此外，我们通过多种定性和定量指标证明了歌词稠密表示的重要性。大量实验表明，所提出的 LTMN 效果显著，能够在输入少量词语作为种子歌词后，自动生成恰当的歌词和富有意义的旋律。

II. 相关工作

过去几年，研究人员一直在探索歌词条件音乐生成任务。ALYSIA [8] 基于随机森林，根据歌词生成旋律。文献 [9]提出了一种用于中国流行歌曲歌词条件旋律生成的编码器-解码器 RNN 序列架构。文献 [10]构建了一个包含对齐信息的大规模配对歌词-旋律数据集，并研究了根据给定的完整歌词生成旋律。作者研究了一种以歌词为条件的端到端深度旋律生成网络，并提出了用于根据歌词生成旋律的条件式长短期记忆网络生成对抗网络（ LSTM-GAN）。文献 [11]的作者通过生成悦耳且有意义、与给定歌词相匹配的旋律，展示了所提出的旋律生成系统的有效性。文献 [12]提出了三分支条件式 LSTM-GAN，并采用 Gumbel-Softmax 技术提升生成质量。此外，文献 [13] 的作者扩展了一个预训练模型，以便更好地捕捉上下文信息。文献 [14]提出了一种可解释的歌词到旋律生成系统。文献 [15]提出了一种用于根据歌词生成旋律的条件混合生成对抗网络模型。

与现有方法不同，本研究探讨如何根据不完整歌词生成恰当的歌词和有意义的旋律。具体而言，首先利用长短期记忆网络，根据给定的种子歌词生成有意义的歌词。以歌词中的每个音节为输入，相应地预测音乐属性（音高、时值和休止符）序列。本文提出采用注意力技术对歌词到旋律生成的序列进行建模。

†贡献相同。

‡Gurunath、Florian 和 Simon 在东京国立情报学研究所（NII）实习期间参与了这项工作。

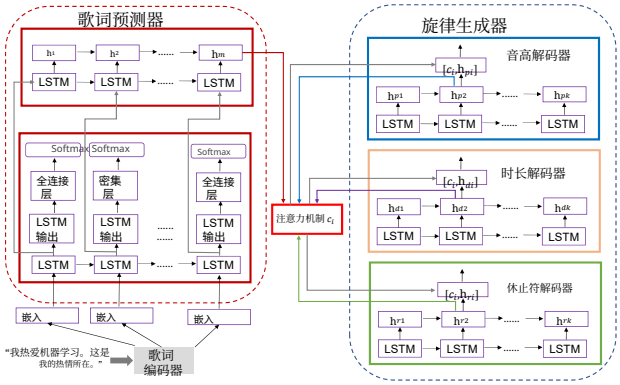


图1：我们的 LTMN 主要由歌词预测模块和旋律生成模块组成。歌词由基于长短期记忆网络的歌词预测器预测。旋律则由分别负责音高、时值和休止符的解码器预测。

III. 所提出的歌词到旋律生成方法

在本研究中，我们设计了一种基于注意力的网络，用于根据不完整歌词生成旋律。模型首先根据种子歌词生成歌词，然后为生成的歌词谱写旋律。我们可以将音乐生成建模为一个概率模型，其中包含两个条件依赖的组成部分：歌词预测器和歌词到旋律生成器，如图1所示。

歌词预测器被建模为一个条件分布，用于根据歌词序列中先前的词元预测下一个歌词词元，该序列表示为 $p(s_t|s_1, s_2, ..., s_{t-1})$。这里，词元是由 $S = \{s_1, s_2, ..., s_T\}$ 给出的歌词音节序列。我们可以学习一个概率分布

$$p(S) = \prod_{t=1}^{T} p(s_t|s_1, s_2, ..., s_{t-1}). \quad (1)$$

每个时刻从学得的概率分布中采样一个词元，以生成完整的歌词序列。

歌词到旋律生成器被建模为一个条件分布模型，给定歌词序列，其旋律序列上的条件分布为 $p(m_1, m_2, .., m_T|s_1, s_2, ..., s_T)$，其中 $M = \{m_1, m_2, .., m_T\}$ 是一个旋律序列，该模型采用序列编码器-解码器模型。编码器是一个长短期记忆网络，在每个时间步依次接收歌词生成模型生成的歌词词元，从而产生编码器隐藏状态 h_t（时间为 t）。编码器还会生成一个可变的上下文向量 C_i，用于编码在旋律解码过程中受到最多注意的歌词。解码器是另一个长短期记忆网络，经过训练后可根据隐藏状态预测时间步 t 的旋律词元。解码器的隐藏状态通过递归方式计算：

$$h_t = g(h_{t-1}, m_{t-1}, C_i). \quad (2)$$

下一个 解码器的词元分布可以定义为 为

$$p(m_t|m_1, m_2, ..., m_{t-1}, C_i) = f(h_t, m_{t-1}, C_i). \quad (3)$$

可以对模型进行联合训练，以最大化似然

$$\max_{\theta} \frac{1}{N} \sum_{i=1}^{N} \log p_\theta(m_i|s_i), \quad (4)$$

其中 θ 是模型参数， m_i， s_i 这一对分别是旋律作曲器的输入和输出。

A. 歌词转向量（*lyric2vec*）

我们训练 Skip-gram 模型，以获取输入歌词文本的稠密表示。这样，我们就能从大型文本语料库中高效地学习高质量的文本词元向量表示 [16]。Skip-gram 模型的目标是：给定一段输入文本中位置 t 处的词元，预测其周围的上下文词元。我们希望模型最大化周围词元的对数概率，计算如下

$$\frac{1}{T} \sum_{t=1}^{T} \sum_{-c \le i \le c, i \ne 0} \log p(s_{t+i}|s_t), \quad (5)$$

其中， c是词元 t^{th} 的上下文长度 s_t。

我们训练 Skip-gram 模型，以获取音节级和词级歌词的嵌入向量，其中 $S = \{s_1, s_2, ..., s_n\}$表示音节词元，$W = \{w_1, w_2, w_3, ..., w_n\}$ 表示词语词元。训练期间，学习率从初始值 0.03 逐渐衰减至 0.0007。我们使用 $c = 7$ 作为上下文窗口长度，负采样分布参数α 设为 0.75。我们训练模型，获得维度为 $v = 50$的音节级和词级嵌入向量。

输入的独热向量表示音节或词，通过线性单元压缩为低维稠密表示，以预测上下文词元。我们分别探讨以下嵌入表示，用于训练歌词预测与旋律生成模型：i）音节嵌入（SE），其中每个音节 s_i 由音节嵌入模型编码为维度为 v的向量。ii）音节与对应词嵌入拼接（SWC），其中音节 s_i 与维度为 v 的对应词嵌入 w_i 拼接。iii）音节与词语向量相加（ASW），即将音节向量与词向量逐元素相加。iv）音节、词语及音节投影词语向量拼接（CSWP）。在 CSWP 中，我们将音节嵌入向量 s_i 投影到对应的词向量 w_i 上，方法是

$$proj_{w_i} s_i = \frac{s_i \cdot w_i}{|w_i|^2} w_i. \quad (6)$$

最后，我们将 s_i、 w_i 和 $proj_{w_i} s_i$ 拼接起来，形成音节 s_i的嵌入向量。

B. 歌词到旋律网络

所提出的 LTMN 是一种序列编码器-解码器模型，用于预测完整歌词并生成相应的旋律。首先，LTMN 将种子歌词作为输入，并开始逐句生成歌词。歌词预测器接收歌词词元序列 S 作为输入，其中每个词元都通过 skip-gram

模型嵌入为一个稠密向量。每个生成的歌词词元在每个时间步都由 RNN编码器编码为一个隐藏向量 h_t 。旋律解码器利用编码后的隐藏向量 h_t 和动态上下文向量 C_t ，在每个时间步生成旋律隐藏状态。每个属性对应的独立解码器接收歌词上下文向量 C_t 和解码器状态，以生成每个音节的音高、时值和休止属性。RNN 中的每个单元均采用门控长短期记忆网络建模 [17]。

注意力向量 c_t 通过计算对齐模型 [18]，将输入歌词与旋律对齐。

$$c_t = \sum_{j=1}^{T} \alpha^{tj} h_t, \tag{7}$$

$$\alpha^{tj} = \frac{exp(e^{tj})}{\sum_{k=1}^{T} exp(e_{tk})}, \tag{8}$$

$$e^{tj} = V_a^T \tanh(W_a \tilde{h}_{t-1} + U_a h_t), \tag{9}$$

其中，参数 V， W_a， U_a 是可学习权重。

神经歌词与旋律创作模型采用端到端方式训练，以最小化交叉熵损失函数的总值。

IV. 实验与结果

A. 数据集

我们在 [10]中创建了实验所用的数据集。在实验中，我们将数据集解析为三种不同的格式：a）音节级：这种格式处于最低层级，将每个音符与对应的音节及其属性配对。b）词级：这种格式将一个词中的所有音符归并在一起，并给出构成该词的每个音节的属性。c）句子级：类似地，这种格式将构成一个音节（或在大多数情况下，构成一行歌词）的所有音符及其对应属性组合在一起。每个音节由离散属性表示：a）音符的音高：我们使用 MIDI音符编号表示音高，其取值范围为 0 到 127 之间的任意整数。b）音符的时值：音符的时值以谱表单位表示。在我们的解析中，它可能取值的完整集合为 [0.125、0.25、0.5、0.75、1、1.5, 2, 4, 8, 16, 32]。以及 c）音符前休止：该值可以取与时值相同的数值，此外也可以为零。

B. 实验设置

所有模型使用的 LSTM 单元数均为 128。旋律解码器模型中的旋律属性被视为词元，每个属性都由 128 维词嵌入表示。使用 Adam优化器将损失函数降至最低，初始学习率为 0.0001，并每 10 个训练轮次线性衰减一次。模型以 32 的批量大小进行训练，以最小化交叉熵损失函数。我们对各种温度值τ 进行的初步实验表明， $\tau \in [0.5, 1.0]$ 是生成

符合数据集分布的歌词-旋律对，用于后续实验。

表 I：种子歌词与生成的完整歌词。

种子歌词	生成歌词
我从未需要过你坚强起来	我从未需要你变得坚强，我从未见过你，你任何时候都…… 我可以……我们能做到……我相信你，我永远都会是那个…… 我爱你，如果你知道，这不过是个女孩，我可以让音乐停下来 在夜里，你能让我有那种感觉；现在我知道你…… 会看到我需要的东西……我感到孤单，我想独自一人 没有一份爱情可以成为我的归宿，我该如何
也许找到我某个地方在你的	也许在你的旅途中，能在某个地方找到我，于是 清晨来临，在我相信 Katy 之前，她走过，那一个就是 我，啪，啪，啪，三年，来摇滚我，假装，让我做自己 那一个，我们飞翔，我会在那里，当我们必须让音乐…… 我们属于彼此，我现在不知道，我知道，我知道，但是我 知道，我知道，我需要，我必须相信，我不会哭，我会 不会让我
我们的世界是绵延数英里的无尽的道路	我们的世界是绵延数英里的无尽道路，一位鼓手走着，我们 总是保持平和，修女快乐地……我从不在意恋爱，而且 在你心里，那就是……我不需要知道是否……我发现 你自己……我永远无法成为你的宝贝，但我不敢相信我竟然…… 不是我一直在等待的那个人……雨中的雨滴一次次让我…… 哭泣……雨中的雨滴……望向外面，我仍然拥有……

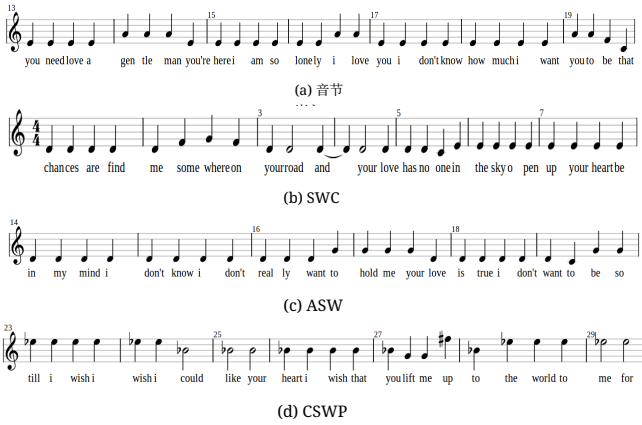


图 2： $\tau = 0.8$ 的 SE、SWC、ASW 和 CSWP 生成歌词-旋律对的乐谱。

C. 生成质量评估

种子歌词和生成的完整歌词示例如表 I 所示。生成结果证明，我们的歌词预测器能够根据作为种子的关键词生成有意义的歌词。

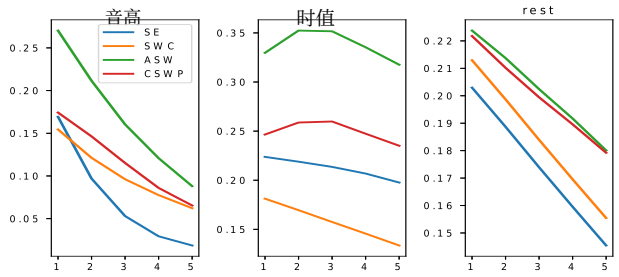


图 3：SE、SWC、ASW 和 CSWP 嵌入表示的音高、时值和休止符 BLEU 分数。

我们使用 BLEU 分数 [19] 评估模型对音高、时值和休止符的预测效果。具体而言，我们计算测试集中的歌词生成的旋律与相应真实旋律之间的 BLEU 分数。BLEU 分数越高，表明生成的旋律与真实旋律之间的相关性越强。不同嵌入表示的结果如图 3 所示。可以看出，在每种音乐属性生成任务中，ASW 和 CSWP 歌词嵌入的表现都优于其他表示。图 2 展示了所提出的 LTMN 生成的歌词与旋律配对的若干乐谱示例。

D. 主观评估

歌词创作和旋律创作是人类的创造性过程，因此客观地、以统计方式评估生成歌词和旋律的质量极具挑战性。因此，我们采用 [10] and [20] 提出的主观评估方法，评估 LTMN 生成的歌词和旋律。在主观评估中，我们向参与者提出以下问题：i）预测的歌词有多大意义？ii）旋律与歌词的契合度如何？参与者按照五分制对给定样本进行评分。我们从所提出的 LTMN、基线和真实数据中，随机选取 5 首生成的完整歌词，以及 5 组歌词-旋律对。

我们遵循 [10] 来创建基线歌词和旋律。基线旋律的属性，即音高、时值和休止符，分别从相应的数据分布中采样。基线歌词则从出现频率最高的音节歌词词汇表中采样。我们分别对生成歌词和歌词-旋律对进行主观评估。

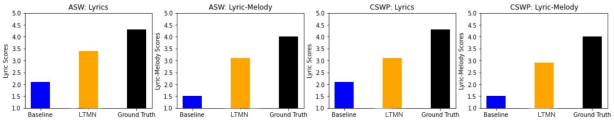


图 4：所提出的 LTMN 在 ASW 和 CSWP 上生成的歌词及歌词-旋律配对的主观评估结果。

11 位至少具备歌词、旋律和音乐创作基础知识的成年人参与了主观评估。 $\tau = 0.6$ ASW 和 CSWP 的主观评估结果如图 4 所示。可以看出，LTMN 的表现接近人类创作。基线的表现不如其他方法，这并不令人意外。我们还发现，ASW 的表现优于 CSWP，这与第 IV 节的自动评估结果一致。从主观评估指标来看，模型生成的歌词和旋律与人类创作之间仍存在差距，这表明可以通过注入先验音乐知识来改进当前模型。

V. 总结

我们提出了一种基于深度注意力机制的对齐网络，用于训练 LTMN，旨在帮助人们发现合适的歌词并生成有意义的旋律。我们在一个大型歌词-旋律平行数据集上训练 lyric2vec 模型，该数据集按音节、词和句子层级进行解析，以提取歌词嵌入。所提出的 LTMN 是一种编码器-解码器序列循环神经网络模型，由针对每种音符属性的歌词预测器和旋律生成器组成，并通过端到端训练和注意力机制学习歌词与旋律之间的关联。此外，实验结果证明，从 ASW 中提取的歌词嵌入优于 CSWP、SE 和 SWC。定性和定量评估表明，所提出的方法能够生成合适的歌词和有意义的旋律。

参考文献

[1] F. Pachet 和 P. Roy，《马尔可夫约束：可控的马尔可夫序列生成》，Constraints，第 16 卷，第 2 期，第 148–172 页，2011 年。[2] F. Pachet、A. Papadopoulos 和 P. Roy，《为结构化音乐生成采样序列变体》，载于 ISMIR，2017 年，第 167–173 页。[3] E. Waite，《在歌曲和故事中生成长期结构》，Magenta Bolg，2016 年。[4] H. Chu、R. Urtasun 和 S. Fidler，《Song from PI：用于流行音乐生成的音乐上合理网络》，arXiv 预印本 arXiv:1611 03477，2016 年。[5] O. 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] H.-W. Dong、W.-Y. Hsiao、L.-C. Yang 和 Y.-H. 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，2018 年。[7] Y. Yu、S. Tang、F. Raposo 和 L. Chen，《用于音乐检索的音频与歌词深度跨模态相关性学习》，ACM Transactions on Multimedia Computing, Communications, and Applications，第 15 卷，第 1 期，第 20 页，2019 年。[8] M. Ackerman 和 D. Loker，《使用 ALYSIA 进行算法作曲》，载于 EvoMUSART，2017 年，第 1–16 页。[9] H. Bao、S. Huang、F. Wei、L. Cui、Y. Wu、C. Tan、S. Piao 和 M. Zhou，《根据歌词创作神经网络旋律》，arXiv:1809.04318，2018 年。[10] Y. Yu、A. Srivastava 和 S. Canales，《用于根据歌词生成旋律的条件式 LSTM-GAN》，arXiv:1908.05551，2019 年。[11] Y. Yu、F. Harsco¨et、S. Canales、G. Reddy、S. Tang 和 J. Jiang，《歌词条件式神经网络旋律生成》，载于 MMM，2020 年，第 709–714 页。[12] A. Srivastava、W. Duan、R. R. Shah、J. Wu、S. Tang、W. Li 和 Y. Yu，《使用三分支条件式 LSTM-GAN 根据歌词生成旋律》，载于 MMM，2022 年，第 569–581 页。[13] Z. Sheng、K. Song、X. Tan、Y. Ren、W. Ye、S. Zhang 和 T. Qin，《SongMASS：通过预训练和对齐约束实现自动歌曲创作》，载于 AAAI，2021 年，第 13 798–13 805 页。[14] W. Duan、Z. Zhang、Y. Yu 和 K. Oyama，《使用离散值对抗训练实现可解释的歌词旋律生成》，载于 ACM MM，2022 年 10 月，第 6973–6975 页。[15] Y. Yu、Z. Zhang、W. Duan、A. Srivastava、R. Shah 和 Y. Ren，《用于根据歌词生成旋律的条件混合生成对抗网络》，Neural Computing and Applications，2022 年 10 月。[16] T. Mikolov、K. Chen、G. Corrado 和 J. Dean，《向量空间中词表示的高效估计》，arXiv 预印本 arXiv:1301 3781，2013 年。[17] K. Cho、B. Van Merri¨enboer、C. Gulcehre、D. Bahdanau、F. Bougares、H. Schwenk 和 Y. Bengio，《用于统计机器翻译的 RNN 编码器-解码器短语表示学习》，arXiv 预印本 arXiv:1406.1078，2014 年。[18] D. Bahdanau、K. Cho 和 Y. Bengio，《通过联合学习对齐和翻译实现神经机器翻译》，arXiv 预印本 arXiv:1409 0473，2014 年。[19] K. Papineni、S. Roukos、T. Ward 和 W.-J. Zhu，《BLEU：一种机器翻译自动评估方法》，载于 ACL，2002 年，第 311–318 页。[20] H.-P. Lee、J.-S. Fang 和 W.-Y. Ma，《iComposer：中国流行音乐自动歌曲创作系统》，载于 NAACL，2019 年，第 84–88 页。