CONTROLLABLE MULTI-STYLE MUSIC GENERATION MODEL BASED ON SIMPLE CONTRASTIVE LEARNING

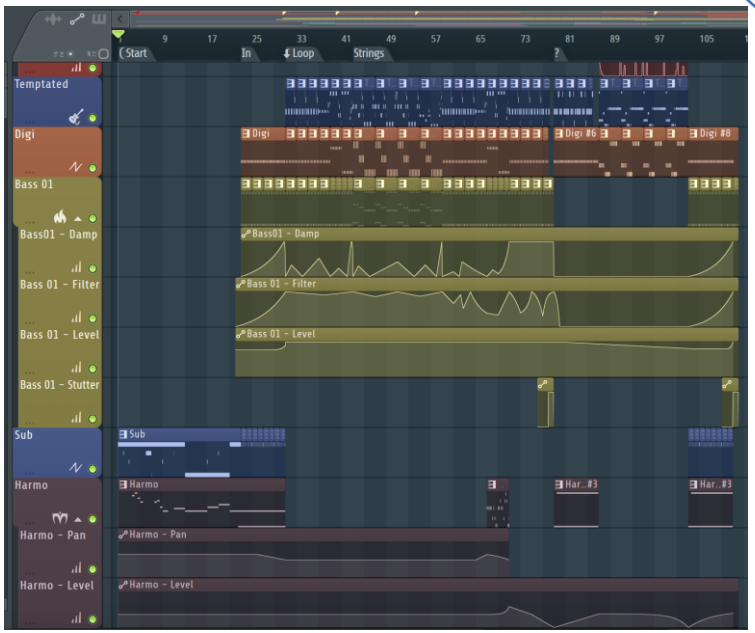
WeiXipin

2022.9.4

2022 how to music?

音乐数据是什么？midi

以前：乐器实录
由录音师傅和乐手、歌手在录音棚
完成音乐制作、成本高

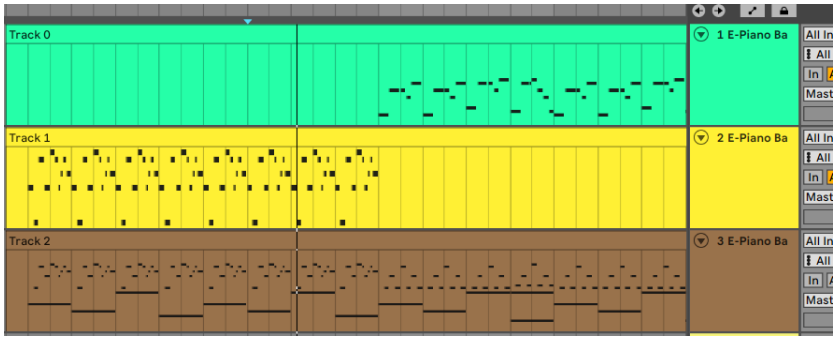


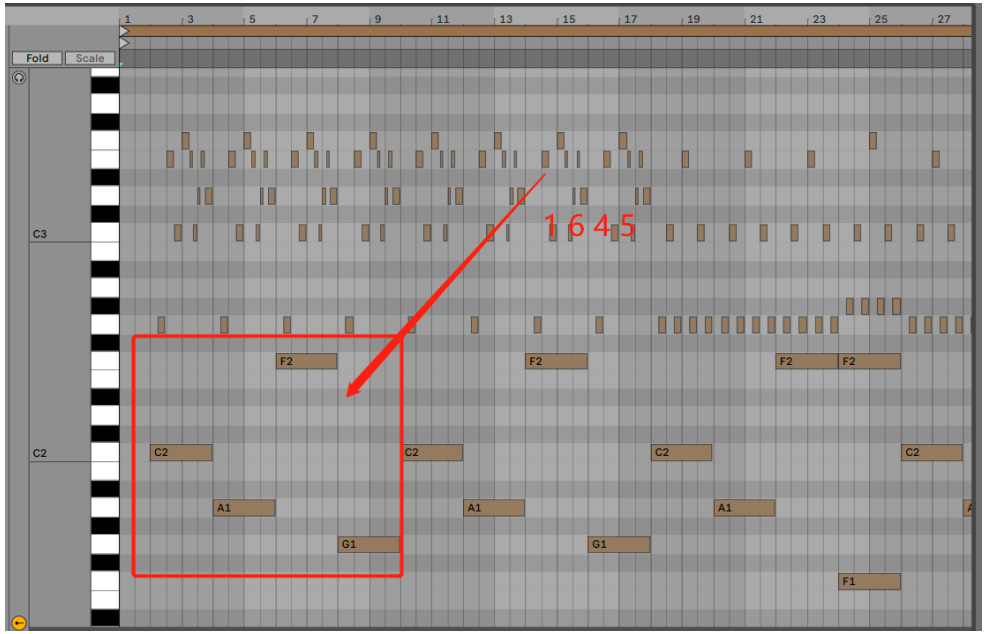
软件音源的迅速发展，人们将大量的音色进行录制，然后封装成软件音源，形成大量的软件音色库。
mid就替代了wav，作为一种可编辑的音乐文件。

近现代：电脑音乐制作
仅仅通过一个编曲/音乐制作人对各种音源音色进行编写，编写多分轨的midi文件，就能够完成一首完整的音乐。

Quick review on music generation

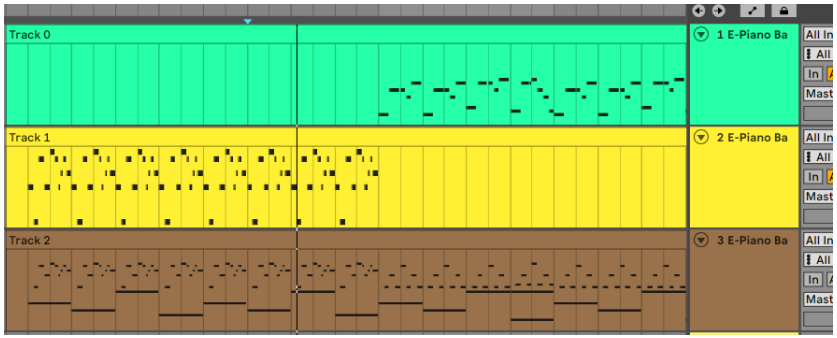
- 音乐生成任务的解决从算法方面来讲，可以分为两大类。
- 第一大类，基于传统乐理知识构建规则的专家算法，这类音乐生成通常由熟知乐理的相关专家来完成，他们将传统的乐理体系总结成计算机可熟知的语言协议或标准，从而实现音乐的生成。遗传算法、 Caos 理论、生成语法、马尔可夫模型。这种类型的创作基于必须以固定顺序遵循的数学指令
- 展示：基于专家算法的我的工程

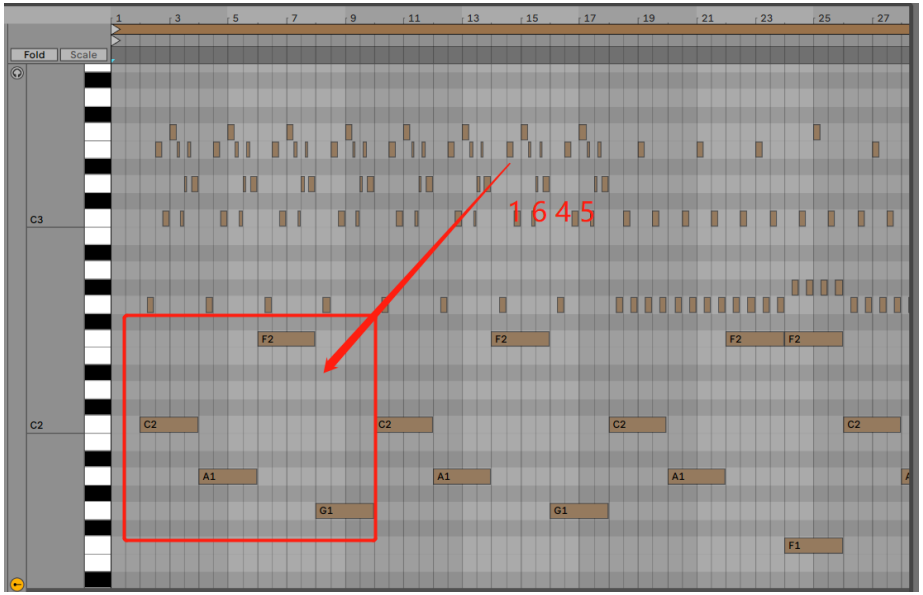




Quick review on music generation

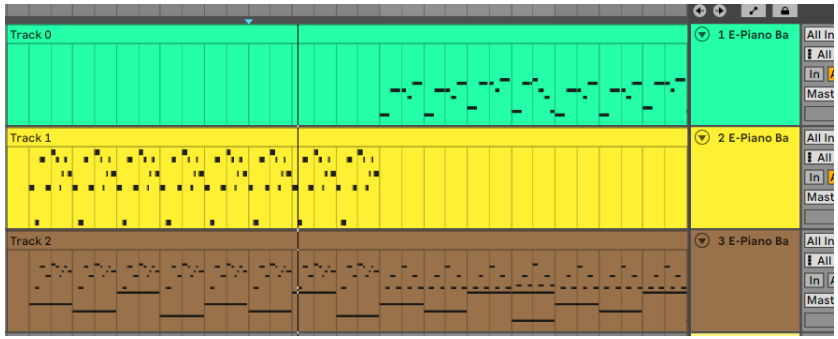
- 第二大类，基于深度学习的音乐生成，MusicVAE、musictransformer、Gan、transformer、Rnn、vae、bert等各种深度学习模型被广泛应用于音乐生成任务。

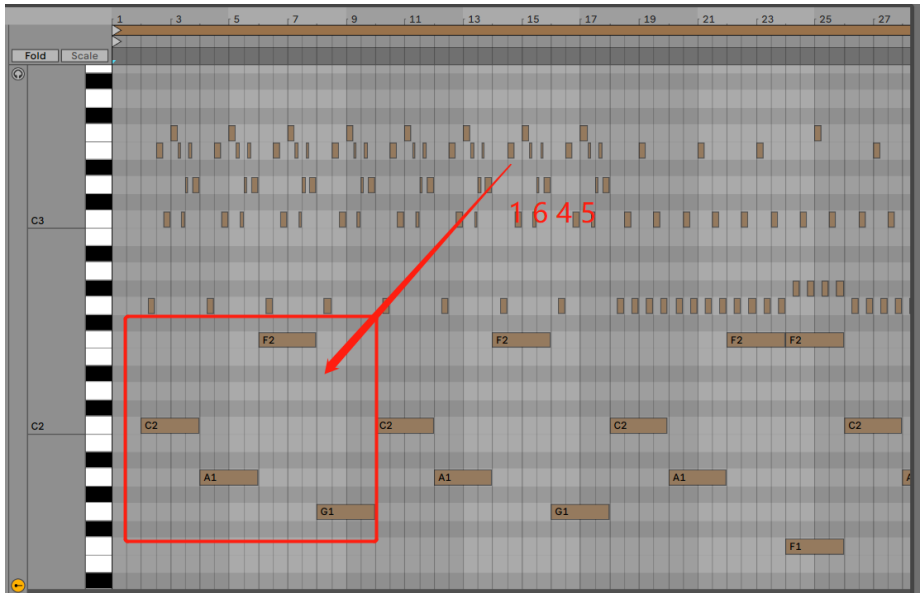




基于深度学习的音乐生成任务

- NLP模型本质：矩阵-输入神经网络-输出矩阵
- 音乐生成任务：音乐特征编码建模(如何构建音乐矩阵)-以矩阵形式输入nlp模型（MusicVAE、musictransformer、Gan、transformer、Rnn、vae、bert）-输出音乐特征解析
- 注意力机制与乐理：musictransformer将注意力机制应用于音乐，并且根据音乐序列对应旋律的周期性和规律性，提出将相对位置信息纳入考量，在音乐生成方面取得了非常不错的效果。





特征建模：音乐特征如何建模成为输入矩阵？

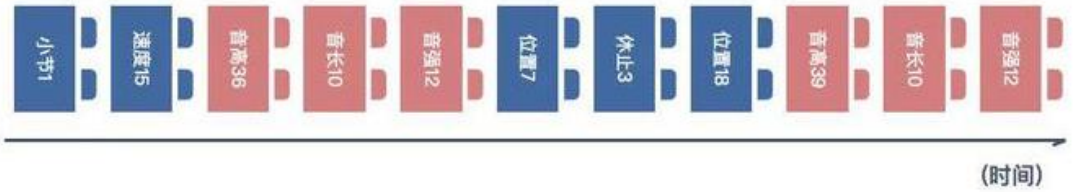
- 三种音乐符号处理方式：Midi-like、REMI、Compound word

	MIDI-like [30]	REMI (this paper)
Note onset	NOTE-ON (0–127)	NOTE-ON (0–127)
Note offset	NOTE-OFF (0–127)	NOTE DURATION (32th note multiples; 1–64)
Time grid	TIME-SHIFT (10–1000ms)	POSITION (16 bins; 1–16) & BAR (1)
Tempo changes	✗	TEMPO (30–209 BPM)
Chord	✗	CHORD (60 types)

Table 1: The commonly-used MIDI-like event representation [23, 30] versus the proposed beat-based one, REMI. In the brackets, we show the ranges of the type of event.

Compound word

- remi、midi-like



- Compound word
- Compound word的引入能对音乐文件进行序列压缩，
- 能够捕捉更长音符间的乐理规律，
- 以更快的速度生成质量更好更长的音乐。
- 更重要的是：它将一些音符特征按照乐理进行了绑定，
- 从而提升了模型的乐理学习

- 一些博客：https://www.163.com/dy/article/H7B42U170511831M.html

不足之处

- 1.乐理规范与音乐创造性兼容（三种midi编码建模方式都在不断提升矩阵与音乐之间的相关性，但是依旧不够完善，这些数学建模依旧会丧失很多音乐特征）
- 比如：音高的设计采用的是简单编码，我们可以考虑引入12平均律对编码 结构进行改进。positon位置编码在音乐市场中已经更新到了tick为最小划分力度，可以实现任意长度拍号的修改，而REMI/CP中最小的划分单位还是1/64拍。

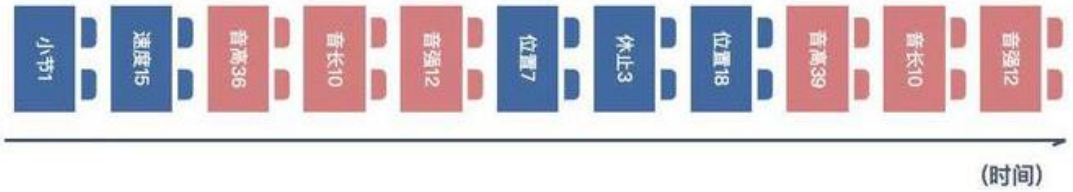
- 2.音乐风格不可控且单一(乐器单一、风格单一)

不足之处：乐理规范与音乐创造性兼容

- 编码建模的改进
- 三种音乐符号处理方式：Midi-like、REMI、Compound word
- Music Transformer、TR autoencoder、Transformer VAE都采用Midi-like的midi Representation，Guitar Transformer、Pop Music TR则是基与REMI的midi Representation，这些模型都没有针对音乐风格特征和乐理规范对midi Representation做更多的设计。大量基于音乐生成模型利用注意力机制确实能有效抓取音符的规律，但却遇到了一个问题，即单纯的注意力机制无法有效在较长的音乐序列中抓到规律，结果只能输出 20~30 秒的旋律，时长增加就难以保证作品的质量，从而音乐作品又会丧失乐理的规范。
- Compound word的引入能对音乐文件进行序列压缩，能够捕捉更长音符间的乐理规律，以更快的速度生成质量更好更长的音乐。

Compound word

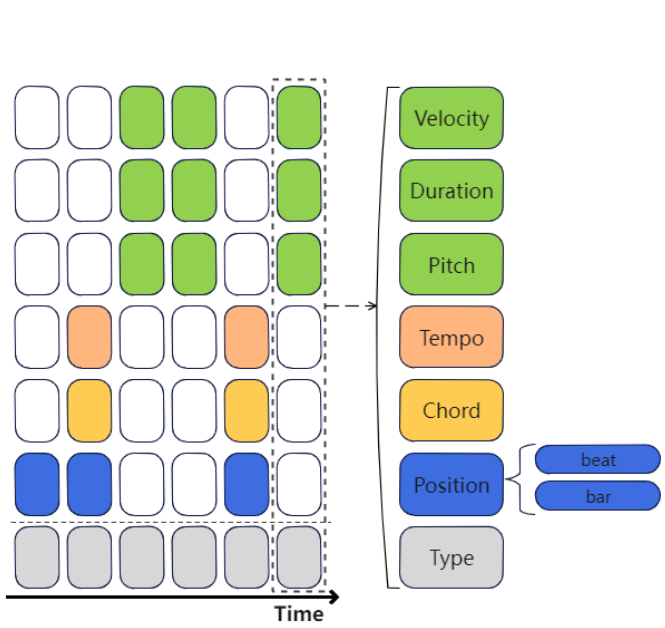
- remi、midi-like

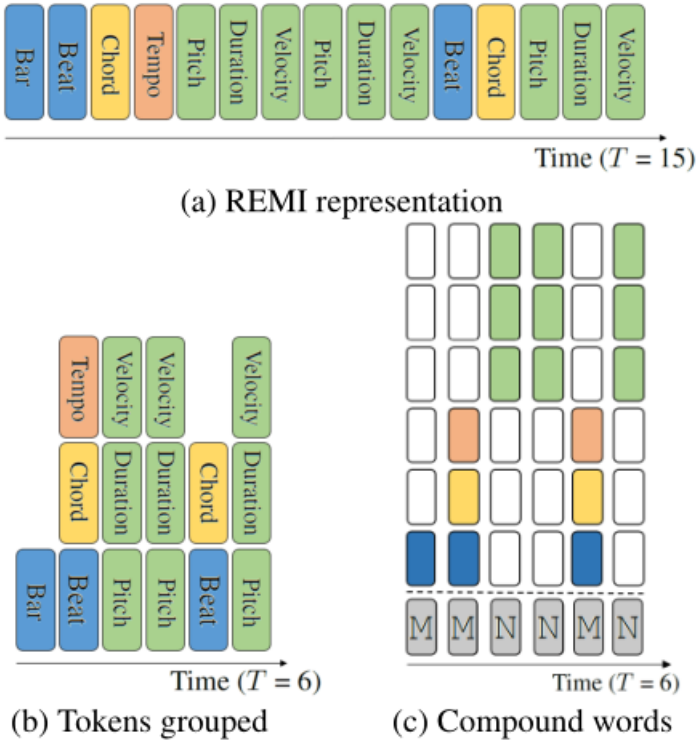


- Compound word
- Compound word的引入能对音乐文件进行序列压缩，
- 能够捕捉更长音符间的乐理规律，
- 以更快的速度生成质量更好更长的音乐。
- 更重要的是：它将一些音符特征按照乐理进行了绑定，
- 从而提升了模型的乐理学习

- 一些博客：https://www.163.com/dy/article/H7B42U170511831M.html

Compound word

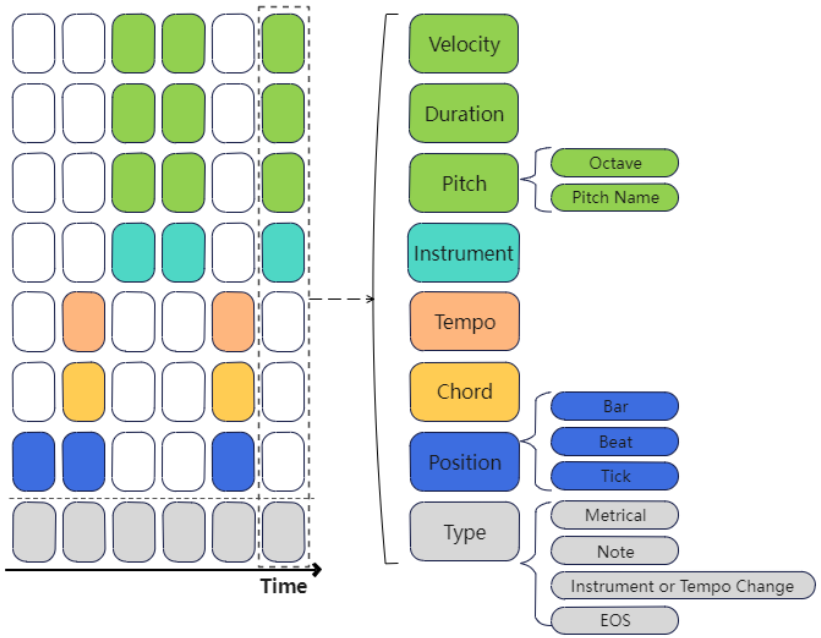




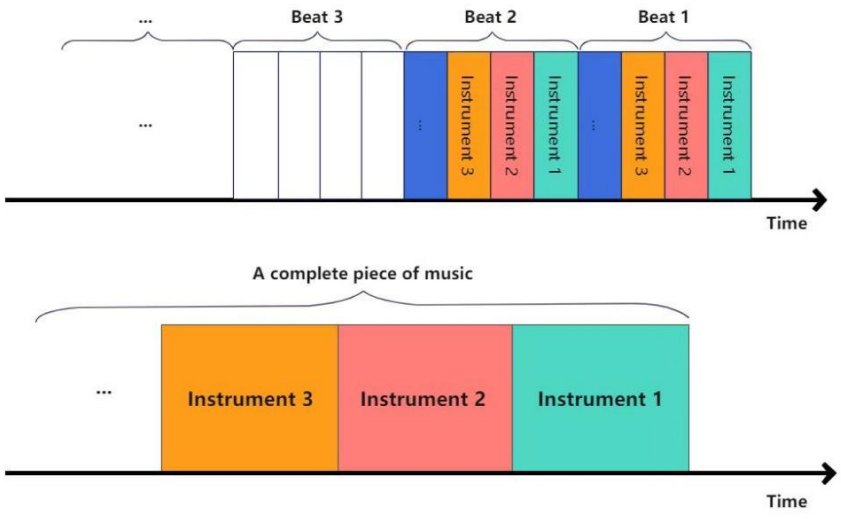
- 缺点：标签不够细，token的设计丧失了一些音乐特征
- 乐器单一、没有风格特征，只是做到了压缩数据，更好的提取音符间的关系。

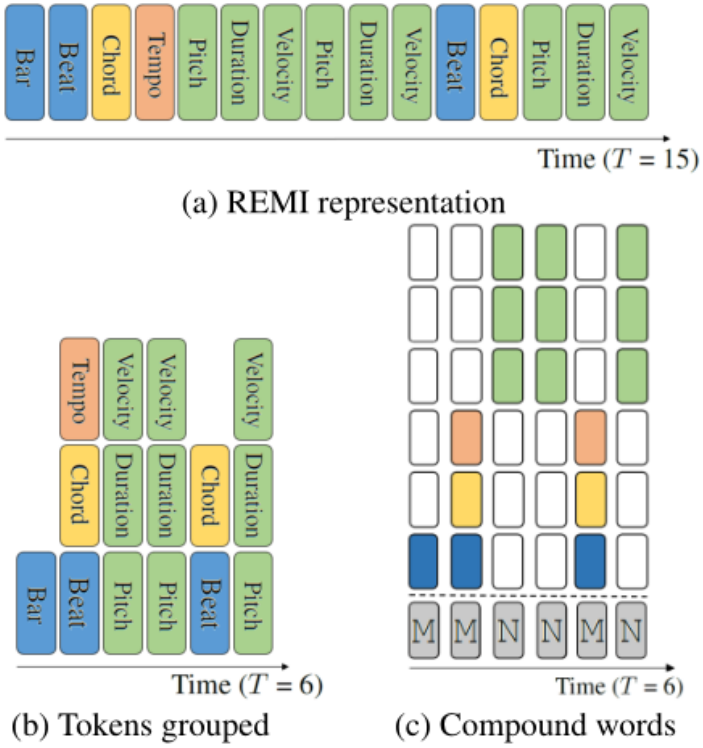
针对多风格的改进

- 一首歌的音乐风格特征通常和音乐的节奏、乐器种类、速度、力度、旋律、和弦关系紧密相关
- 重新设计标签



新的插入方式


...
Beat 3
Beat 2
Beat 1
Instrument 3
Instrument 2
Instrument 1
Time
A complete piece of music
Instrument 3
Instrument 2
Instrument 1
Time

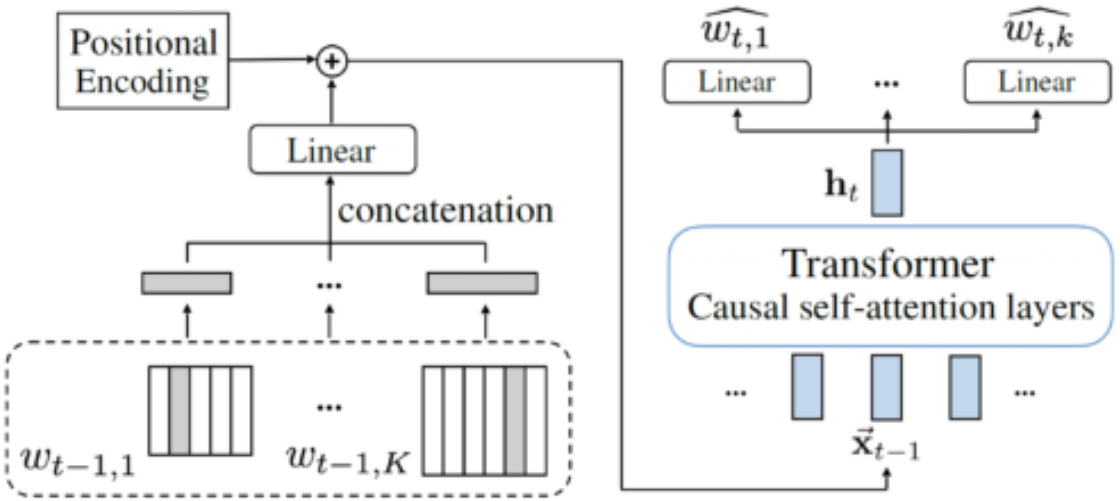

Bar
Beat
Chord
Tempo
Pitch
Duration
Velocity
Time (T = 15)
(a) REMI representation
Time (T = 6)
(b) Tokens grouped
M
N
Time (T = 6)
(c) Compound words

不足之处：音乐风格不可控且单一

- 数据集缺少针对性分类：目前的音乐生成模型Pop Music Transformer对于单一音乐风格就行模型训练，Guitar Transformer针对单一乐器进行训练，Music Transformer、Musicvae等模型训练都赖于数据集，并没有对数据集进行分类整理，这些都使得每次所生成的音乐流派风格具有很大的随机性，很难进行人为控制。没有风格特征。
- 利用对比学习来提升模型对风格特征的理解。使得模型生成的音乐风格根据不同的输入具有可控性。

backbone

$$
\begin{aligned}
\mathbf{h}_t &= \text{Self-attn}\left(\vec{\mathbf{x}}_{t-1}\right) , \\
\widehat{f}_t &= \text{Sample}_{\mathcal{F}}\left(\text{softmax}(\mathbf{W}_{\mathcal{F}}\mathbf{h}_t)\right) , \\
\mathbf{h}_t^{\text{out}} &= \mathbf{W}_{\text{out}}\left[\mathbf{h}_t \oplus \text{Embedding}_{\mathcal{F}}(\widehat{f}_t)\right] , \\
\widehat{w_{t,k}} &= \text{Sample}_k\left(\text{softmax}(\mathbf{W}_k\mathbf{h}_t^{\text{out}})\right) , \; k = 1, ..., K ,
\end{aligned}
$$



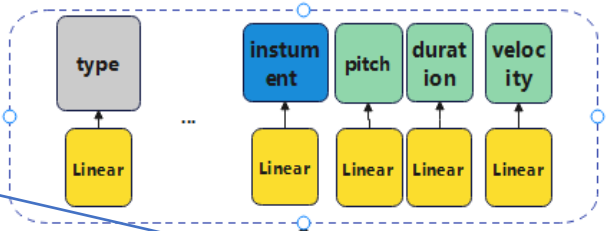
$$
\begin{aligned}
\mathbf{p}_{t,k} &= \text{Embedding}_k(w_{t,k}) , \; k = 1, ..., K , \\
\mathbf{q}_t &= \text{Embedding}_{\mathcal{F}}(f_t) , \\
\mathbf{x}_t &= \mathbf{W}_{\text{in}}\left[\mathbf{p}_{t,1} \oplus ... \oplus \mathbf{p}_{t,K} \oplus \mathbf{q}_t\right] , \\
\vec{\mathbf{x}}_t &= \text{Positional Encoding}(\mathbf{x}_t) ,
\end{aligned}
$$

对比学习

- simcse-风格特征

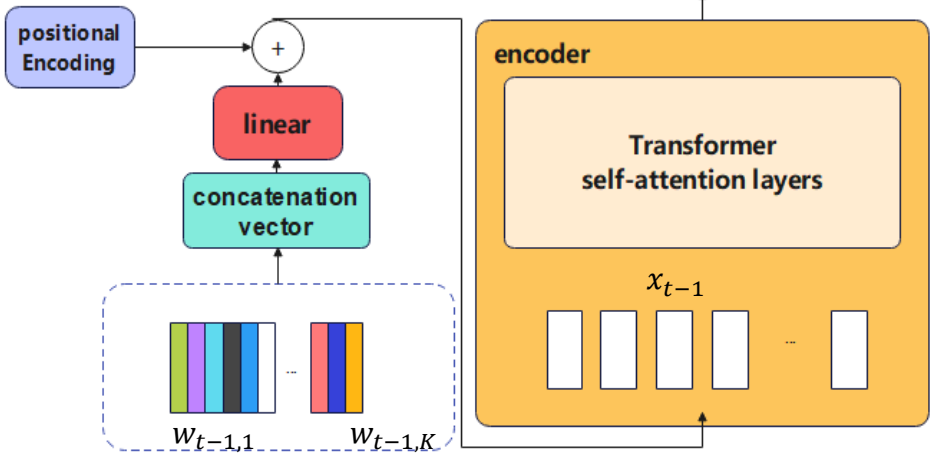
	样本1	样本2	样本3	样本4
样本1	1	0	1	1
样本2	0	1	1	0
样本3	0	0	1	0
样本4	1	0	0	1

$$\mathbf{h}_t = \text{Self-attn}\left(\vec{\mathbf{x}}_{t-1}\right),$$
$$\widehat{f}_t = \text{Sample}_{\mathcal{F}}\left(\text{softmax}(\mathbf{W}_{\mathcal{F}}\mathbf{h}_t)\right),$$
$$\mathbf{h}_t^{\text{out}} = \mathbf{W}_{\text{out}}\left[\mathbf{h}_t \oplus \text{Embedding}_{\mathcal{F}}(\widehat{f}_t)\right],$$
$$\widehat{w_{t,k}} = \text{Sample}_k\left(\text{softmax}(\mathbf{W}_k\mathbf{h}_t^{\text{out}})\right),\ k = 1, ..., K,$$

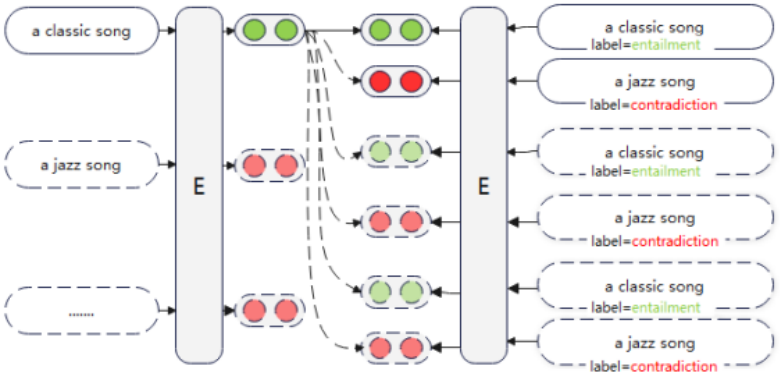


$$log(cCL) = -\log \frac{e^{\text{sim}\left(\mathbf{h}_i, \mathbf{h}_i^+\right)/\tau}}{\sum_{j=1}^{N}\left(e^{\text{sim}\left(\mathbf{h}_i, \mathbf{h}_j^+\right)/\tau} + e^{\text{sim}\left(\mathbf{h}_i, \mathbf{h}_j^-\right)/\tau}\right)} \quad (13)$$

$$loss = loss(CompoundWord) + \alpha * loss(CL) \quad (14)$$



(b) Supervised Simple Contrastive Learning

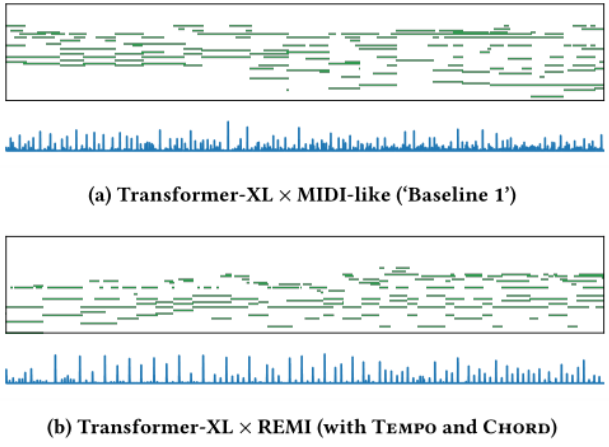


$$\mathbf{p}_{t,k} = \text{Embedding}_k(w_{t,k}),\ k = 1, ..., K,$$
$$\mathbf{q}_t = \text{Embedding}_{\mathcal{F}}(f_t),$$
$$\mathbf{x}_t = \mathbf{W}_{\text{in}}\left[\mathbf{p}_{t,1} \oplus ... \oplus \mathbf{p}_{t,K} \oplus \mathbf{q}_t\right],$$
$$\vec{\mathbf{x}}_t = \text{Positional Encoding}(\mathbf{x}_t),$$

评价指标：缺乏统一的评价指标

指标模版：客观评价（针对不同的task创新点，每个模型都有自己的评价方式）
+主观评价（问卷打分）

- pop music transformer： remi的提出，最大的特点是引入了位置信息



(a) Transformer-XL × MIDI-like ('Baseline 1')

(b) Transformer-XL × REMI (with TEMPO and CHORD)

Figure 1: Examples of piano rolls and 'downbeat probability curves' (cf. Section 5.3) for music generated by an adaptation of the state-of-the-art model Music Transformer [23], and the proposed model. We can see clearer presence of regularly-spaced downbeats in the probability curve of (b).

客观： bar-beat-节奏信息的追踪和对比

主观： 问卷打分

评价指标：缺乏统一的评价指标

指标模版：客观评价（针对不同的task创新点，每个模型都有自己的评价方式）+主观评价（问卷打分）

compound word 复合词

Task	Representation + model@loss	Training time	GPU memory	Inference (/song) time (sec)	Inference (/song) tokens (#)	Matchness melody	Matchness chord
Conditional	Training data	—	—	—	—	0.755	0.838
	Training data (randomized)	—	—	—	—	0.049	0.239
	REMI + XL@0.44	3 days	4 GB	88.4	4,782	0.872	0.785
	REMI + XL@0.27	7 days	4 GB	91.5	4,890	0.866	0.800
	REMI + linear@0.50	3 days	17 GB	48.9	4,327	0.779	0.709
	CP + linear@0.27	0.6 days	10 GB	29.2	18,200	0.829	0.733
Unconditional	REMI + XL@0.50	3 days	4 GB	139.9	7,680	—	—
	CP + linear@0.25	1.3 days	9.5 GB	19.8	9,546	—	—

Table 4: Quantitative evaluation result of different models. REMI+XL represents a re-implementation of the state-of-the-art Pop Music Transformer (Huang and Yang 2020), while CP+linear stands for the proposed CP Transformer.

客观：gpu占用量和生成速度、旋律的相关性（规定音符在一个八度内的越多越高）、和弦相关性(色度向量)

主观：问卷打分

评价指标：缺乏统一的评价指标

- MIDI-VAE 风格迁移：三个独立的风格评估分类器

客观：三个独立的风格评估分类器：
音高、力度、乐器
然后将三个独立的音乐风格评估综合加权打分

	Pitch		**I**nstrument		**S**tyle		**V**elocity	
	Train	Test	Train	Test	Train	Test	Train	Test
CvJ	0.90	0.85	0.99	0.87	0.98	**0.92**	0.008	**0.029**
CvP	0.96	**0.88**	0.99	**0.89**	0.96	0.91	0.017	0.036
JvP	0.88	0.80	0.99	0.86	0.94	0.69	0.043	0.048
BvM	0.91	0.75	0.99	0.82	0.94	0.74	0.010	0.033

Table 2. Train and test performance of our final models. The velocity column shows MSE loss values, whereas the rest are accuracies.

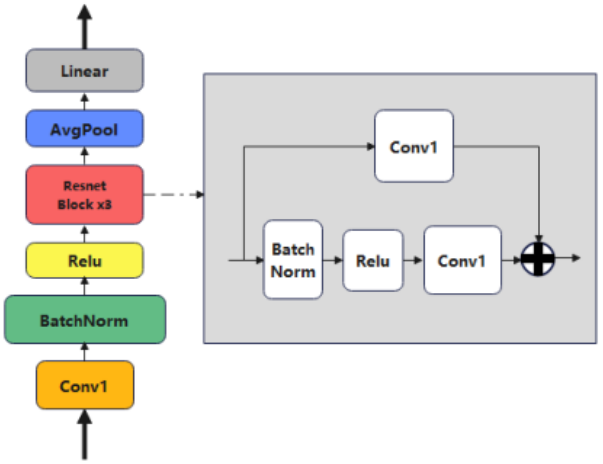
	Pitch		Velocity		Instrument	
	Bf.	Af.	Bf.	Af.	Bf.	Af.
CvJ Test	0.77	0.66	0.67	0.57	0.90	0.20
CvP Test	0.77	0.67	0.71	0.60	0.91	0.27
JvP Test	0.65	0.63	0.67	0.64	0.67	0.55
BvM Test	0.55	0.47	0.60	0.49	0.64	0.47

Table 4. Average before and after classifier accuracies for all classifiers (pitch/instrument/velocity) for the test set.

评价指标

- 本文借鉴MIDI-VAE的评价方式：1.训练一个三分类文本任务来评价生成音乐类型的正确性，利用cnn和renet将音乐特征全部提取出来，包括音高、力度、乐器，还有更多的音乐位置特征信息。

pitch	velocity	duration	instrument	tempo	chord	position	type



- 2.主观评价：问卷打分（生成结果在https://music.163.com/album?id=148363059&userid=1854909036）

展示视频-模型生成midi-midi傻瓜式导入移动端工程

Future work

- 完善对比实验组
- 1、扩展风格种类，现在针对：Jazz、Classic、Pop三种风格的音乐进行实验
- 2、第一组对比实验：compound word 原文的复现（最原始的token）与现在的改进token后的模型
- 3、第二组对比实验：使用有监督对比学习与不使用对比学习的实验组