北京交通大學

《Jukebox: A Generative Model for Music》阅读报告

姓名：吉言

学号：17211046

学院：电子信息工程学院

任课教师：李磊

模块负责老师：陈一帅

时间：2020 年 5 月

一．文献信息

文献名称：《Jukebox: A Generative Model for Music》

文献作者：Prafulla Dhariwal；Heewoo Jun；Christine Payne; Jong Wook Kim; Alec Radford; Ilya Sutskever- OpenAI

发表途径：OpenAI

发表时间：January 1962

二．问题意义

自动音乐的产生的研究已经有悠久的历史。目前已经有生成钢琴乐谱的算法，可生成歌手声音的数字声码器，以及为各种乐器产生音色的合成器等等。每一个都捕捉音乐产生的特定方面：旋律，作曲，音色和人声演唱。但是，目前仍然无法使用一个系统就能够完成所有的音乐生成功能。且使用符号方法对音乐进行的建模虽然容易，但是其设计结果很难捕捉人声，也无法很好地表现出音乐中的手法力度表现力等。本文所设计的 Jukebox-自动点唱机，是一种基于神经网络的机器学习模型框架，当原始音乐在一定范围内的类型和音乐风格的情况下，该框架可以生成音乐（包括基本歌曲）。

三．思路方法

1.准备工作

作者旨在设计一个能够从音色到整体连贯性上捕捉各色音乐的模型。使用自动编码器能够解决音乐时长问题提升学习效率。该自动编码器通过丢弃一些感知上不相关的信息位，将原始音频压缩到较低维度的空间。然后我们令我们的训练模型在这个低维度空间生成音频，并向上采样回到原始音频空间。而在原始音频中，要让模型学会解决多样性以及超长距离结构等问题。

首先对上图的原始音频进行 CNN 编码如下图：

上图：压缩音频：每秒 344 个样本。

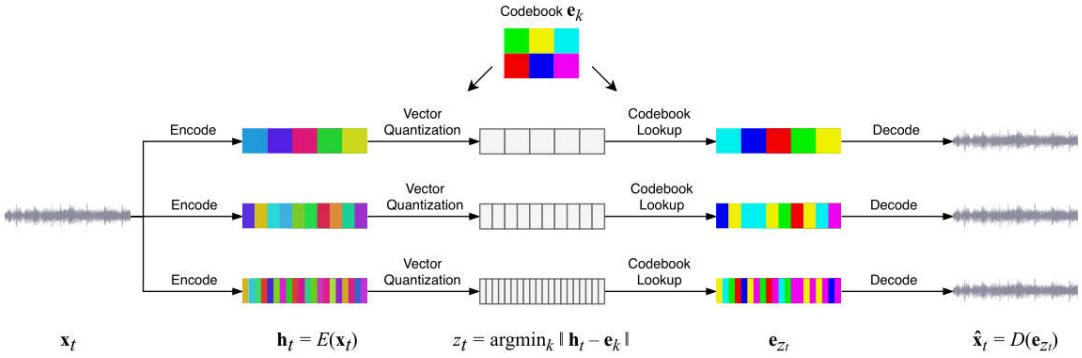
然后产生新颖的模式-通过以歌词为条件的受训转化器，结果见下图：

最后使用转化器升频和使用 CNN 解码，见下图

2.对音乐的 VQ-VAE 方法

作者使用 VQ-VAE 将音乐压缩为离散代码。

分层 VQ-VAE 方法可以从几套乐器中生成简短的乐器片段。但是由于使用和自回归解码器匹配的连续的编码器，会遭受层次崩溃的困扰。一个称为 VQ-VAE-2 的简化变体，通过仅使用前馈编码器和解码器避免了这些问题，并且在生成高保真图像时，显示出令人印象深刻的结果。如下图所示：



为了缓解 VQ-VAE 带来的码本崩溃，作者使用随机重启机制；为了最大限度利用层次，作者也使用了单独的编码器；还添加了一个光谱损失来使得模型能够轻松重构更高的频率。光谱表达式如下：

$$\mathcal{L}_{\text{spec}} = \| |\text{STFT}(\mathbf{x})| - |\text{STFT}(\hat{\mathbf{x}})| \|_2$$

4. 音乐优先和上行采样器

训练 VQ-VAE 之后，我们需要学习压缩空间上的先验 p（z）以生成样本。我们将先验模型分解为：

$$p(\mathbf{z}) = p(\mathbf{z}^{\text{top}}, \mathbf{z}^{\text{middle}}, \mathbf{z}^{\text{bottom}}) \quad (5)$$

$$= p(\mathbf{z}^{\text{top}})p(\mathbf{z}^{\text{middle}}|\mathbf{z}^{\text{top}})p(\mathbf{z}^{\text{bottom}}|\mathbf{z}^{\text{middle}}, \mathbf{z}^{\text{top}}) \quad (6)$$

作者从起始价、流派和时间调节上对模型进行了其它训练，接着让模型适应歌曲的歌词来提供一些上下文信息。从简单的试探法开始，逐步获得歌词的精确对齐。为了学习歌词，作者采用了编码解码模型（Encoder-decoder model），添加了一个编码器来生成歌词的表示形式，并添加了注意层，这些注意层使用了音乐解码器中的查询，来关注歌词编码器中的键和值。训练后，模型将学习更精确的对齐方式。

为了减少训练歌词条件模型所需的计算，我们使用预先训练的无条件顶级先验作为解码器。

在我们完成 VQ-VAE 的训练，上采样，和顶级先验之后，我们可以用他们去采样新的歌曲。我们采样的方式有三种：祖先采样，窗口采样和灌注采样。

四．实验结果

为了训练该模型，作者挑选出 120 万首歌曲，组成一个新数据集（其中 60 万为英语），并与来自 LyricWiki 的歌词和元数据配对。元数据包括艺术家，专辑类型和歌曲的年份，以及与每首歌曲相关的常见状态或播放列表关键字。我们对 32 位、44.1 kHz 的原始音频进行训练，并通过随机下混右声道和左声道来执行数据增强，以产生单声道音频。

作者在 VQ-VAE 中使用三个级别，分别将 44kHz 原始音频压缩为 8x，32x 和 128x 三个级别，每个级别的码本大小为 2048。这种下采样会损失许多音频细节，并且随着我们进一步降低级别，听起来会有明显的噪声。但是，它保留有关音频的音高，音色和音量的基本信息。

每个 VQ-VAE 级别独立编码输入。底层编码产生最高质量的重构，而顶层编码仅保留基本的音乐信息从而达到压缩的目的。

为了生成新颖的歌曲，一系列的转换器会从上到下生成代码，然后，下层的解码器可以将它们转换为原始音频。

最终作者得到了 Jukebox 模型。作者通过手动评估生成的样本的连贯性，音乐性，多样性和新颖性来完成对音乐的评估。

接下来使用转换器生成代码。然后我们训练现有模型，其目标是学习由 VQ-VAE 编码的音乐代码的分布，并在此压缩离散空间中生成音乐。像 VQ-VAE 一样，我们具有三个优

先先验级：生成最高压缩码的顶级优先级，以及两个上采样优先级，用来生成上述情况的较少压缩的代码。

最高级别的先验模型，对音乐远距离结构进行了建模，从此级别解码的样本具有较低的音频质量，但捕获了诸如唱歌和旋律之类的高级语义。中间和底部的上采样，先验添加了诸如音色之类的本地音乐结构，从而大大改善了音频质量。

五．启发思考

该学习模型旨在捕捉音乐的产生形式，从而生成由模型自己产生的音乐。但是作者自己也承认，该模型还是远远无法替代音乐人的地位，只能在针对于那些对音乐感兴趣但没有经过正规培训的人。但是这个模型依然具有相当大的进步意义，已经能够产生数分钟长的乐曲，并具有自然声音识别的歌声。

我在运行源代码的过程中，因对 anaconda 的使用不甚熟悉，对 python3 的学习也仅限于皮毛，因此最后还是没能够成功产生一段简短的音乐，不得不说这是这次研学的一个极大的遗憾。但是我还是从这次研学的过程获益良多，希望以后还有机会继续这个项目的学习与研究，也感谢老师的指导。