堆叠卷积自编码器对数梅尔频谱图重构能力的研究

Anastasia Natsiou都柏林理工大学都柏林，爱尔兰
anastasia.natsiou@tudublin.ie

Luca Longo都柏林理工大学都柏林，爱尔兰
luca.longo@tudublin.ie

Se´an O'Leary都柏林理工大学都柏林，爱尔兰
sean.oleary@tudublin.ie

arXiv:2301.07665v1 [cs.SD] 18 Jan 2023

摘要——**在音频处理应用中，基于高层表示生成表现力声音的需求日益增长。这些表示可用于操控音色，并影响创意乐器音符的合成。神经网络等现代算法推动了基于乐器音色压缩的表现力合成器的发展。无监督深度学习方法可以通过训练网络学习从波形或频谱图到低维表示的映射，从而实现音频压缩。本研究探讨了使用堆叠卷积自编码器对多种乐器在单一音高下的时频音频表示进行压缩。进一步探索超参数和正则化技术，以提升初始设计的性能。通过无监督方式，网络能够根据潜在表示重构单声部谐波声音。此外，我们提出一种评估指标，用于衡量原始样本与重构样本之间的相似度。评估用于声音合成的深度生成模型是一项颇具挑战的任务。我们的方法基于生成频率的准确性，因为它是感知谐波声音的一项重要度量指标。本研究有望推动未来使用神经自编码器进行音频压缩的实验。**

索引词—**对数梅尔频谱图重建、自编码器、机器学习。**

I. 引言

内存有限以及对数据快速传输的要求，促使研究人员和工程师探索降低数据维度的方法。此前已有计算方法被提出以解决这一问题。主成分分析（*PCA*） [1]是一种算法，它通过提取最重要的信息作为一组新的正交变量，将每个样本线性投影到一个维度更低的空间。

近来，深度学习方法的兴起展现了将数据非线性投影到低维空间的能力。无监督网络，如自编码器（AE）[2]或生成对抗网络（GANs） [3] ，已证明能够从相对庞大且复杂的数据集中提取信息并映射到潜在空间，效果颇具前景。此类架构最初用于从图像 [4] 或视频 [5]中提取有意义的模式。然而，也有少数研究也提出将其用于语音 [6] 乃至音乐压缩任务 [7]。

在这项工作中，我们旨在研究堆叠卷积神经自编码器，以降低谐波声音时频表示的维度。自编码器（AE）是一种由编码器和解码器组成的神经网络架构。编码器将输入表示投影到维度较低的空间，而解码器则尝试将该表示还原至原始维度。

卷积神经网络属于人工神经网络（ANN）的一个子类，旨在基于局部模式检测从输入样本中提取特征。该网络的特点是节点之间存在空间局部连接，并使用一个扫过输入矩阵的特征图。池化层等其他技术还可以进一步降低数据的维度。

本文介绍了一种卷积自编码器的设计，并探讨了利用正则化方法对其进行改进。通过使用特定音高下的多种乐器，本文探索了音色压缩。该网络尝试将单声部谐波声音的对数梅尔频谱图映射到低维空间。压缩后的高层表示用于乐器音色合成。

本文结构安排如下：第二节回顾了利用自编码器压缩音频表示的相关工作。第三节介绍自编码器、卷积网络、正则化技术和梅尔频谱图的相关细节，第四节则展示所开展的实验。第五节介绍所采用的评估方法，并提出一种新颖的评估指标。最后，第六节展示结果并进行讨论，第七节则简要总结全文。

II. 相关工作

神经自编码器已用于此前旨在从音频表示中提取有意义信息的研究。在这类研究中，大多数神经网络都用于压缩幅度短时傅里叶变换帧 [7]–[11]。最早的尝试之一见于 [8] ，其中

浅层前馈神经自编码器成功将原始表示压缩了 25%。研究人员还采用了去噪技术，进一步对自编码器进行正则化。在此前工作的基础上，Colonel 等人[7] 开发了一个四层多层感知机，使用 Adam 优化器和加性偏置，并采用 L1 和 L2 进行权重正则化。在上述两项研究中，他们通过计算原始幅度谱图帧与生成帧之间的均方误差 (MSE) 来评估各自的方法。 [9] 也应用了前馈自编码器，将 2049 点快速傅里叶变换 (FFT) 帧转换为包含 8 个值的潜在空间。为提升网络性能，他们采用了数据增强技术，使用幅度谱的一阶和二阶差分，并额外加入梅尔频率倒谱系数。研究人员通过对生成帧应用短时傅里叶逆变换 (ISTFT) 来评估该架构，同时还提出了一种确定性的相位重建方法。此外，Colonel 等人 [11] 还研究了使用自编码器进行音色合成时的激活函数。

在 [10] 中，对幅度短时傅里叶变换帧的生成进行了更深入的研究。研究将浅层和深层前馈网络，以及循环自编码器和变分自编码器与 PCA 进行了比较。研究表明，只有深层自编码器和长短期记忆网络能够充分重建原始样本。模型通过均方误差进行了主观评估，也通过对质量评分进行统计分析进行了客观评估。当潜在空间未发生变化时，使用原始相位通过短时傅里叶逆变换合成波形；当需要估计相位时，则使用 Griffin-Lim 算法 [12] 。

我们的方法主要基于 [13] 中开发的基线自编码器。在这项工作中，研究人员开发了神经自编码器，以从音频片段中提取有意义的信息。他们尝试了不同的输入表示，例如原始音频、短时傅里叶变换的实部和虚部，或功率谱的对数幅度部分。模型通过原始频谱图与生成频谱图之间的均方误差进行了客观评估，也使用 彩虹谱图进行了视觉比较；彩虹谱图是以瞬时频率着色的频谱图。

III. 方法论

本研究旨在了解堆叠卷积自编码器重建对数梅尔频谱图的能力。图1展示了网络的设计：基于卷积网络的编码器旨在将乐器声音压缩到较低维度的空间中，而镜像解码器则尝试根据这一高层表示重建样本。我们建议读者在继续阅读本文的同时，聆听所开展实验合成的声音。音频文件 [1] 的波形是使用 Griffith Lim 算法 [14] 以及正弦信号重建方法 [15]合成的。

[1]https://anastasianat57.github.io/StackedConvolutionalAutoencoders

为改进卷积自编码器的设计，我们采用正则化技术来防止过拟合并提升网络性能。在本研究中，我们评估自编码器在声音合成方面的有效性，并通过使用预计能够校准模型的附加技术来衡量可能的改进。本节将详细介绍自编码器、卷积神经网络及其组成部分，回顾正则化方法，并介绍梅尔频谱图。

A. 自编码器

自编码器是一种神经网络架构，由编码器和解码器组成，二者 [2]以无监督方式共同训练。编码器尝试通过非线性过程提取有意义的信息，从而降低原始样本的维度。随后，解码器旨在根据中间表示重建原始样本。这个生成的中间表示称为嵌入或潜在表示。

对于前馈单层模型，编码器使用非线性激活函数 $f(.)$ 将输入向量 $x \in \mathbb{R}^d$ 映射为编码 $z \in \mathbb{R}^e$，其中 $d > e$

$$y = f(Wx + b) \tag{1}$$

其中，$W \in \mathbb{R}^{(e \times d)}$ 表示神经元之间连接的权重，$b \in \mathbb{R}^e$ 表示偏置项。解码器使用类似的方法将 z 映射回重建后的 $\widehat{x} \in \mathbb{R}^d$

$$\widehat{x} = f(W_{out}y + b_{out}) \tag{2}$$

其中 $W_{out} \in \mathbb{R}^{(d \times e)}$ 和 $b_{out} \in \mathbb{R}^d$。

为了将初始架构扩展为深层神经自编码器，我们采用了类似的方法。编码器将神经元的输入 x_i 映射为输出表示 x_j，用于 $i,\ j < n$如下:

$$x \to x_1 \to x_2 \to ... \to x_n \tag{3}$$

而解码器则以一般方式反转上述过程，如下所示:

$$x_n \to x_{n+1} \to x_{n+2} \to ... \to \widehat{x} \tag{4}$$

自编码器通过代价函数进行训练，该函数旨在最小化原始样本与生成样本之间的误差，并使用反向传播更新 W 和 b 的值。自编码器所使用的激活函数和代价函数因应用场景和训练数据而异。

B. 卷积网络

卷积神经网络（CNNs）构成了多层感知机的一种正则化版本。它们的正则化基于这样一个事实：前馈神经网络研究整个样本中的关系，而卷积网络则在局部搜索模式。其架构

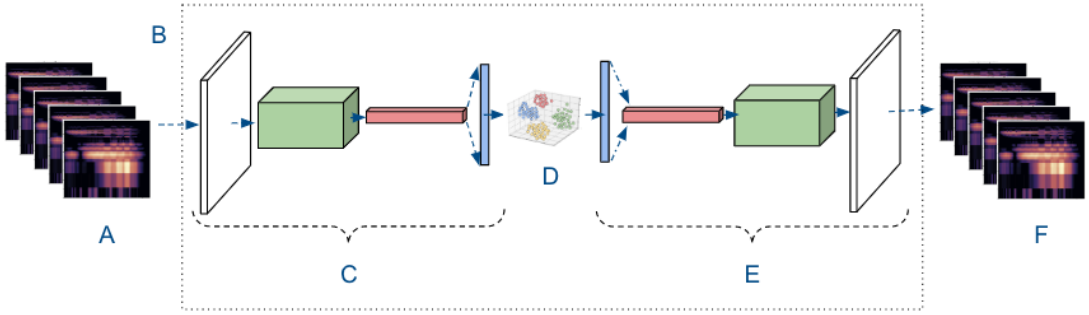

图1. 用于重构对数梅尔频谱图的堆叠卷积自编码器架构示意图。(A) 原始对数梅尔频谱图。(B) 堆叠卷积自编码器。(C) 基于三个卷积层的编码器。(D) 多个样本的潜在表示投影。(E) 基于对称编码器的解码器。(F) 重构后的对数梅尔频谱图。

由三个主要部分组成：卷积层、池化层和全连接层。

1) 卷积层：卷积层是卷积神经网络中最重要的模块。它由一组尺寸小于训练样本的滤波器和一组可学习参数组成。滤波器应用于输入数据，并生成激活图。卷积层由若干参数定义，例如滤波器的数量、在输入数据上移动的步数（步幅），以及零填充，即在输入矩阵外围补零，使滤波器能够适配输入尺寸的过程。

2）池化：可选择在每个卷积层之后，接入一个池化层来处理卷积后的输出。池化是一种降维过程，即在卷积后的输出上应用最大值或平均值滤波矩阵。虽然池化会丢失一些信息，但这种技术可以提高效率、降低复杂度，并降低过拟合风险。

3）全连接层：与卷积层不同，全连接层将每个输出节点直接连接到前一层。该层可根据从卷积层提取的特征执行分类，并调整输出的维度。

C. 正则化技术

深度神经网络最突出的问题之一是容易对训练数据集过拟合。过拟合模型是一种深度学习模型，它将噪声也考虑在内，从而刻意模仿训练样本。这类模型包含了过多的参数，因此在预测新数据时准确性较低。为防止过拟合，人们提出了许多技术，包括随机失活 [16], *L1* 和 *L2* 正则化 [17], 数据增强[18], 或 提前停止 [19]。

1) 随机失活：一种技术，在每次迭代中，网络会随机选择一些节点，并将其排除在学习过程之外 [16]。因此，每次迭代都会由不同的一组节点来预测输出。随机失活也可以视为一种集成技术，因为每个学习步骤都会创建多个子网络。

2) L1 和 *L2* 正则化：这两种方法都会在网络的代价函数中添加一个正则化项。这个附加项会为损失函数增加一个残差，从而减少过拟合[17]。L1 正则化（也称为 *Lasso*）使用权重参数的最小绝对偏差（LAD）来更新代价函数，如下所示：

$$Cost function = Loss + \frac{\lambda}{2m}\sum \|W\| \tag{5}$$

其中，λ 对应正则化参数，m对应参数数量。采用 L2 正则化时，会将权重矩阵的最小二乘误差（LSE）加入代价函数：

$$Cost function = Loss + \frac{\lambda}{2m}\sum \|W\|^2 \tag{6}$$

正则化可以仅应用于权重 W（核正则化），也可以仅应用于偏置项 b （偏置正则化），或应用于输出层 $y = Wx + b$（活动正则化）。

3) 数据增强：通过对原始样本进行交替处理，尝试增加训练数据集。这些额外数据会为网络引入额外噪声，可用于防止过拟合。在图像处理中，可以通过翻转、缩放或平移原始图像来实现数据增强 [18]；而在音频处理中，常见技术包括噪声注入、时间平移或速度交替 [20]。

4) 提前停止：使用验证数据集在每个训练轮次后计算损失函数。如果网络在训练集上的性能持续提升，而在验证集上的性能没有改善，则模型开始过拟合。提前停止用于防止网络模仿训练数据的模式 [19]。耐心值表示在训练停止前，性能未进一步改善的训练轮次数。

解码器. 梅尔频谱图

频谱图是通过对音频样本中

重叠的时间片段应用傅里叶变换得到的时频表示。梅尔频谱图是基于人类感知对频谱图进行压缩后的形式。梅尔尺度的建立基于这样的观察：人耳在较低频率处具有更高的分辨率 [21]。赫兹频率 f 与梅尔尺度 mel （mel 源自“旋律”）之间的对应关系如式 7 所示。

$$mel = 2595log_{10}(1 + \frac{f}{700}) \tag{7}$$

梅尔频谱图以压缩形式捕捉了声音最重要的特性。因此，这种时频表示已被证明有益于深度学习应用，因为它降低了内存和算力需求。

IV. 实验

A. 数据集

在所进行的实验中，我们使用了 NSynth 数据集[2], 的一个子样本。该数据集包含时长为四秒的单音音符。这个子样本由 3750 个样本组成，涵盖多种乐器：吉他、贝斯、铜管乐器、合成器、键盘、长笛、管风琴、槌击乐器、人声、簧片乐器和弦乐器，且音高相同。这些声音可以是原声的、电子的或合成的，并可根据其力度或音质归入不同类别。

数据预处理包括使用 690 个样本的布莱克曼窗、大小为 1024 的快速傅里叶变换窗和 128 个梅尔滤波器组计算对数梅尔频谱图。随后将对数梅尔频谱图归一化，使其转换到范围 $[0, 1]$。数据集按 80/10/10 的比例划分为训练集、验证集和测试集。

B. 模型

所提出的方法展示了一种堆叠式卷积自编码器，其编码器和解码器呈镜像结构，如图1所示。这两个组件均由三个二维卷积层构成，卷积核大小为 4、步幅为 2，并采用同尺寸填充。经过实验研究，卷积层采用双曲正切作为激活函数，而输出层则采用 Softmax 函数来生成对数梅尔频谱图。研究还进一步考察了卷积核维度、滤波器和池化技术。实验也测试了其他正则化方法，包括提前停止、随机失活、L1 和 L2 卷积核正则化，以及活动正则化。网络使用 ADAM 优化器 [22] 进行训练，初始学习率为 0.001，采用均方误差损失函数，批次大小为 64。提前停止的耐心值上限设为 10，以避免在训练过程中浪费资源。实验使用 TensorFlow 库[3]在 Tesla P100 GPU 上进行。

[2]https://magenta.tensorflow.org/datasets/nsynth [3] https://www.tensorflow.org/

V. 评估

为评估生成网络的有效性，人们提出了许多方法。在这些实验中，采用原始频谱图与生成频谱图之间的均方根误差（RMSE）进行评估。另一项用于衡量生成频谱图准确性的度量指标是结构相似性指数（SSIM），它也被证明是比较两幅频谱图的有效初步指标。然而，这两项客观指标都无法准确反映频谱图的质量。频谱图中任何一个数值的差异都会产生相同的均方误差，无论该数值处于什么位置，但合成声音可能并不相同。此外，如图 2 所示，频谱图中的大多数数值均为零，因此即使是较小的误差，也可能体现出两幅频谱图之间的差异。

为了评估生成的频谱图，我们设计了一种度量指标，用于突出谐波在音色合成中的重要性。我们的方法基于峰值检测算法，逐帧估算原始频谱图（*OrigFreq*）和生成频谱图（*GenFreq*）的频率。随后，将相同的频率（*IdFreq*）视为检索到的频率；这些频率与相关频率的差值在 $\pm$3% 的阈值范围内。此阈值可确保检测到的峰值对应于同一个谐波。此外，为了消除可能检测到的噪声，峰值振幅低于峰值振幅 30 dB 的谐波不予考虑。

随后，可利用原始频率、生成频率和相同频率的计算结果，通过精确率-召回率框架评估预测样本 [23]。精确率的计算方式为：在预测频率中，由相同频率表示的真正例值所占的比例，如公式 8 所示。

$$Precision = \frac{IdFreq}{GenFreq} \tag{8}$$

召回率是原始频谱图中估计的频率数量里相同频率的数量。召回率也称为灵敏度，如公式 9 所示。

$$Recall = \frac{IdFreq}{OrigFreq} \tag{9}$$

精确率和召回率是用于计算检索值与相关值之间相关性的度量指标。这两种误差随后可以根据公式 10 合并为一个称为 *F1* 分数的单一度量值。F1 分数体现了精确率和召回率的调和平均数，能够为整体系统提供更准确的度量指标。

$$F1_score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{10}$$

精确率、召回率和 F1 分数的取值范围为 0 到 1，数值越高表示性能越好。精确率高而召回率低，说明生成的样本中

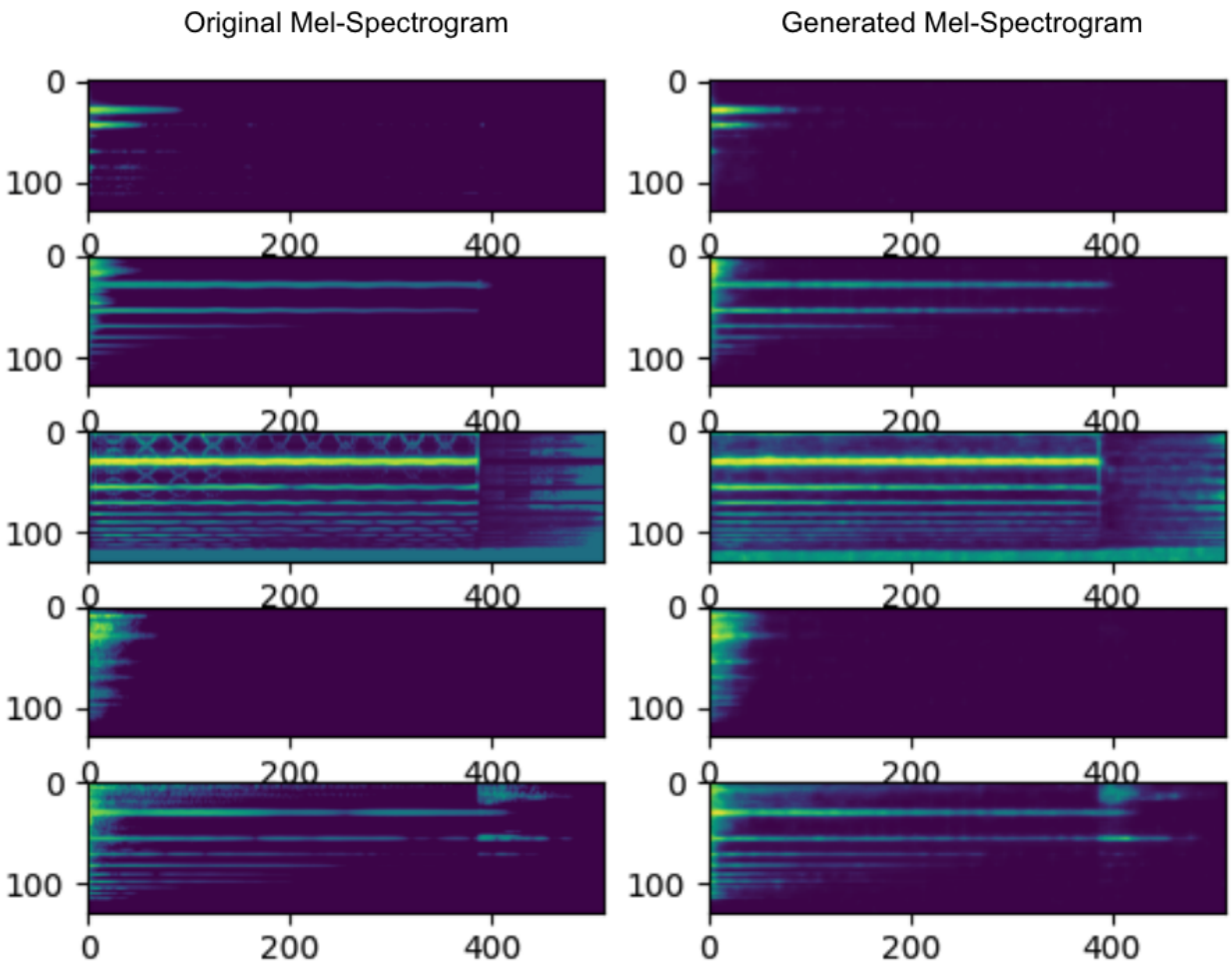

图 2. 原始与生成的对数梅尔频谱图

检测到的谐波较少；而精确率低、召回率高，则说明合成频谱图中存在更多噪声。

VI. 结果与讨论

本研究的主要目标是深入探讨如何使用堆叠卷积自编码器合成音色。额外的参数化和正则化方法旨在尽量减小原始音频样本与预测音频样本之间的误差。本段的各个部分分别给出一组实验参数的结果。首先，我们介绍基线自编码器，然后讨论采用正则化方法、池化、不同潜在空间维度以及不含全连接层的卷积网络进行的实验。

A. 基线

基线设计采用神经自编码器，由三个卷积层、一个最大池化层和一个全连接层组成。编码器生成维度为 8192 的潜在表示，网络训练时除提前停止外不采用其他正则化方法。该原始和生成的对数梅尔频谱图如图 2 所示。可以看出，重建后的频谱图能够捕捉原始样本的一般形态 [4]。

B. 正则化

为提升基线网络的性能，我们研究了其他改进方法。表 I 展示了采用不同正则化技术的实验。具体而言，我们应用正则化方法，以提高自编码器重建对数梅尔频谱图的效果。随机失活的概率设为 0.3，并分别仅应用于编码器（E）、仅应用于解码器（D），或同时应用于编码器和解码器（ED）。此外，我们还进行了核正则化器（KR）和活动正则化器（AR）的实验，分别在编码器、解码器或两个网络上采用 L1 或 L2 正则化。

结果表明，随机失活并未提升网络的性能。具体而言，仅在解码器端使用随机失活时，均方根误差与基线相同，但 F1 分数较低。然而，召回率提高而

[4]https://anastasianat57.github.io/StackedConvolutionalAutoencoders

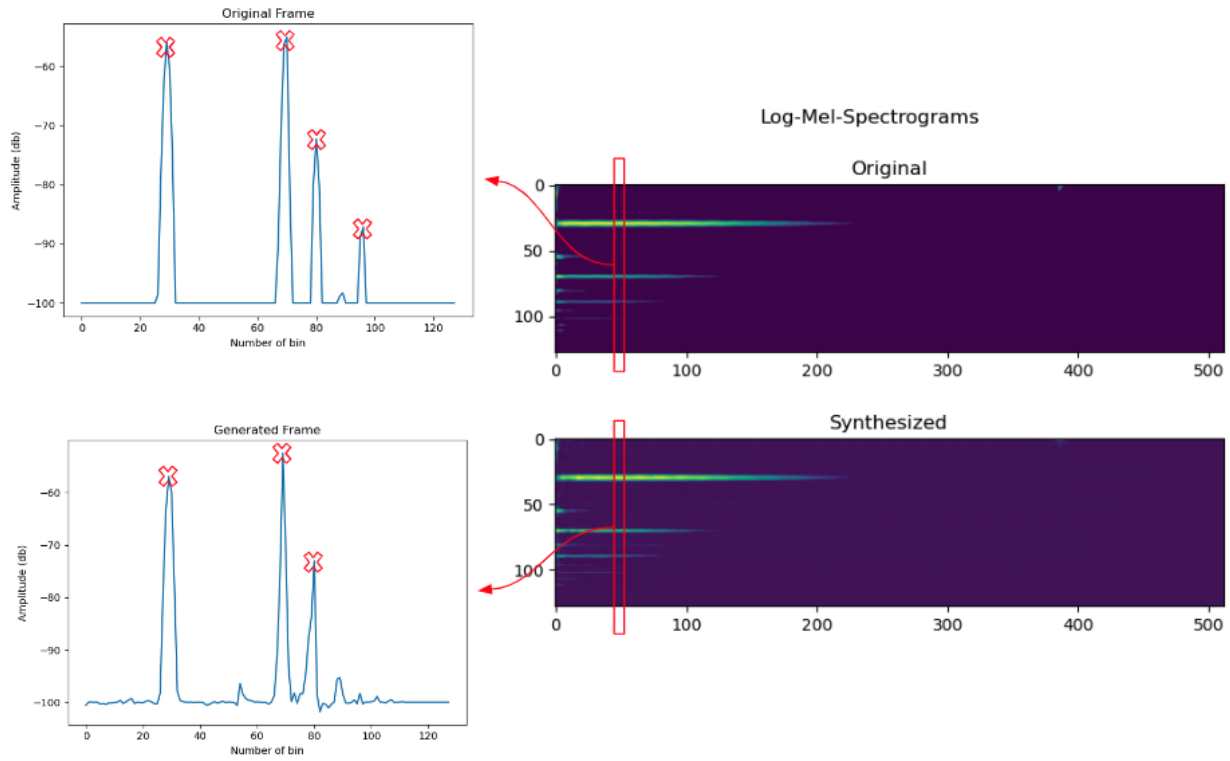

图 3. 原始频谱图与生成频谱图之间的峰值检测

实验	RMSE	SSIM	召回率	精确率	F1
无正则化	0.087	0.733	0.777	0.786	0.772
E端丢弃	0.097	0.686	0.707	0.797	0.732
D端丢弃	0.087	0.640	**0.832**	0.738	0.769
E、D端丢弃	0.102	0.548	0.765	0.753	0.740
KR L1 E	0.102	0.631	0.762	0.752	0.743
KR L1 D	0.079	0.757	0.820	0.804	0.804
KR L1 编码器和解码器	0.080	0.761	0.829	0.794	0.804
KR L2 编码器	0.086	0.722	0.804	0.791	0.790
核正则化器 L2 解码器	**0.077**	**0.782**	0.819	0.829	**0.817**
核正则化器 L2 编码器和解码器	0.079	0.766	0.793	**0.830**	0.801
活动正则化器 L1 编码器	0.085	0.778	0.788	0.826	0.798
活动正则化器 L1 解码器	0.244	0.214	0	0	0
活动正则化器 L1 编码器和解码器	0.244	0.215	0	0	0
活动正则化器 L2 编码器	0.162	0.395	0.541	0.697	0.552
活动正则化器 L2 解码器	0.244	0.177	0	0	0
活动正则化器 L2 编码器和解码器	0.238	0.210	0	0	0

表 I 正则化技术：比较未采用正则化方法的网络、在编码器 (E)、解码器 (D) 或编码器和解码器 (ED) 上采用随机失活的网络，以及采用核正则化器 (KR) 和活动正则化器 (AR) 的网络

精确率降低，表明模型在生成样本中产生了更多噪声。此外，在编码器端或编码器和解码器两端使用随机失活，显著降低了网络的性能。

采用核正则化器作为正则化方法，可以发现性能有所提升。在解码器上采用 L1 或 L2 正则化函数，均可获得最准确的结果。具体而言，在解码器上采用 L2 核正则化器，得到最高的 F1 分数和结构相似性指数 (SSIM)，以及最低的均方根误差 (RMSE)。这些结果具有较高的精确率和召回率，因此生成的频谱图最具潜力。在编码器上，或同时在编码器和解码器上采用核正则化器，也能带来小幅提升。最后，活动正则化器削弱了神经网络的能力，导致出现非谐波结果。在大多数情况下，生成的声音都只是简单的噪声。

总体而言，堆叠卷积自编码器能够从低维表示中合成乐器音色。在某些情况下，L1 或 L2 卷积核正则化等正则化技术能够提升网络的性能，使其生成更准确的频谱图。然而，随机失活等其他正则化技术并未带来任何改进；而在更极端的情况下，例如使用活动正则化器时，解码器会生成随机噪声。

C. 池化

下一组实验基于卷积网络的参数化。更具体地说，我们对池化层进行了研究。

实验	RMSE	SSIM	召回率	精确率	F1
最大池化	**0.087**	**0.733**	0.777	**0.786**	**0.772**
平均池化	0.089	0.708	**0.787**	0.770	0.770
无池化	0.106	0.676	0.692	0.732	0.692

表 II 最大池化、平均池化和无池化卷积网络的比较

实验	RMSE	SSIM	召回率	精确率	F1
稠密层	0.087	0.733	0.777	0.786	0.772
无稠密层	**0.063**	**0.861**	**0.856**	**0.859**	**0.855**

表 IV 卷积块末尾不含全连接层的网络

实验	RMSE	SSIM	召回率	精确率	F1
潜在维度 8192	0.087	0.733	0.777	0.786	0.772
潜在维度 4096	**0.077**	0.796	**0.813**	0.828	0.814
潜在维度 2048	0.078	**0.798**	0.810	**0.837**	**0.817**

表 III 潜在空间进一步压缩的实验

具体而言，我们研究了池化层。最初，在编码器和解码器的每个卷积层之后都采用了最大池化层或平均池化层。最后，我们还研究了不使用池化层的实验。为了使不含池化层的网络与含池化层的网络具有相同的潜在维度，我们额外添加了三个卷积层。表 II 展示了基于不同池化技术的实验结果。根据均方根误差、结构相似性指数和 F1 分数，采用最大池化或平均池化的网络性能相近。尽管 F1 分数几乎相同，平均池化的召回率更高、精确率更低，这表明样本中存在额外噪声。最后，不使用池化层的网络生成的合成声音质量较低。根据上述结果，我们可以得出结论：采用粗略池化方法（例如最大池化）的堆叠卷积自编码器，能够从乐器有音高声音的压缩低维空间中生成更准确的音频时频表示。

D. 潜在空间的维度

为评估神经自编码器的压缩能力，对潜在空间维度进行了实验研究。通过生成维度为 4096 和 2048 的潜在表示，研究了进一步下采样的效果。结果见表 III。根据各项度量指标，压缩程度更高的表示能够生成更准确的对数梅尔频谱图。尽管预期较低的潜在维度会降低网络性能，但实验表明，较小的潜在空间能够合成高次泛音分布更加平滑的频谱图，从而产生更自然的声音。

E. 不使用全连接层

在最后一组实验中，我们尝试移除自编码器的全连接层。全连接层对于调整潜在空间的维度并进行进一步压缩是必要的。然而，表 IV 表明，不含全连接层的卷积自编码器生成的频谱图质量更好。这可能是因为网络保留了原始样本的空间信息。

F. 讨论

从上述实验可以看出，堆叠卷积自编码器能够以足够的精度压缩和重构声音的频谱表示。如图 2 所示，对数梅尔频谱图中的大多数频带值均为零，因此整个表示是否都可视为实际编码的信息尚有争议。然而，听觉具有很高的灵敏度，频谱图的细微变化就可能合成出显著不同的声音。因此，网络必须极其精确地处理非零值的位置和振幅。

与最先进水平相比 [13]，，我们的声音具有相位相干性，听起来更加悦耳。然而，在整个 NSynth 数据集上根据音高对自编码器进行训练，虽然可以丰富生成声音的多样性，但也会带来额外的复杂性，从而降低网络的性能。此外，使用 Wavenet [24] 作为声码器可能无法保留波形中的相位连续性，因此会生成更多带噪声的乐器声音。

VII. 结论

此前已有研究利用深度生成模型从压缩表示中合成音色。这些研究大多根据短时傅里叶变换幅度谱的帧生成声音。为此，研究人员提出了循环自编码器或全连接自编码器。本研究探讨了基于听觉感知的频谱图重建所采用的堆叠卷积自编码器，并对自编码器的参数化技术和正则化方法进行了实验。此外，我们还提出了一种评估方法，用于计算单音和谐波音乐声音中预测频率的准确度。实验结果表明，压缩至原始数据量 3% 的潜在表示能够以足够的准确度合成音色。此外，在应用于网络的多种正则化技术中，有些技术相较于基线自编码器有所改进，而另一些则限制了神经网络的能力。预计本研究将

推动未来有关音色压缩以及生成的谐波声音评估的研究。

未来的研究将拓展自编码器的能力，纳入音高和乐器方面更多样化的声音。我们将研究通过利用额外的时间或频谱信息对网络进行条件化处理，以进一步压缩表示。此外，预计在编码器部分末端加入门控循环单元或长短期记忆网络等时序层，将影响网络的性能。最后，未来的工作将利用更多感知属性，对潜在空间开展实验。继 Esling 等人 [25] 和 Roche 等人 [26],的先前研究之后，我们将围绕音色描述符展开调查，以改善自编码器生成的潜在表示结构，并提高合成音频的准确性。可采用多种数值或可视化解释技术，对生成的潜在空间进行解读和评估 [27], [28]。

致谢

本研究由爱尔兰科学基金会及其机器学习研究培训中心（18/CRT/6183）资助。

参考文献

[1] H. Abdi 和 L. J. Williams，“主成分分析：主成分分析”，《*Wiley*跨学科计算统计评论》，第2卷，第433–459页，2010年7月。[2] G. E. Hinton 和 R. R. Salakhutdinov，“利用神经网络降低数据维度”，《科学》，第313卷，第504–507页，2006年7月。[3] I. Goodfellow、J. Pouget-Abadie、M. Mirza、B. Xu、D. Warde-Farley、S. Ozair、A. Courville 和 Y. Bengio，“生成对抗网络”，《*ACM*通讯》，第63卷，第139–144页，2020年10月。[4] A. Radford、L. Metz 和 S. Chintala，“利用深度卷积生成对抗网络进行无监督表示学习”，2016年1月。arXiv:1511.06434 [cs]。[5] E. L. Denton 和 v. Birodkar，“从视频中无监督学习解耦表示”，载于《神经信息处理系统进展》，第30卷，Curran Associates, Inc.，2017年。[6] W.-N. Hsu、Y. Zhang 和 J. Glass，“学习用于语音生成与转换的潜在表示”，*Interspeech*，第1273–1277页，2017年。[7] J. Colonel、C. Curro 和 S. Keene，“改进用于音乐合成的神经网络自编码器”，纽约，第6页，2017年。[8] A. Sarroff 和 M. A. Casey，“使用自编码神经网络进行音乐音频合成”，国际音乐信息检索学会会议论文集（*ISMIR2014*），2014年10月。[9] J. Colonel、C. Curro 和 S. Keene，“作为音乐音频合成器的自编码神经网络”，载于国际数字音频效果会议（*DAFx*）论文集，葡萄牙阿威罗。，2018年。[10] F. Roche、T. Hueber、S. Limier 和 L. Girin，“音乐声音建模的自编码器：线性、浅层、深层、循环和变分模型的比较”，2019年5月。编号：arXiv:1806.04096 arXiv:1806.04096 [cs, eess]。[11] J. T. Colonel 和 S. Keene，“为实时音色插值与合成而调节自编码器潜在空间”，载于*2020*年国际神经网络联合会议（*IJCNN*），第1–7页，IEEE，2020年。[12] D. Griffin 和 Jae Lim，“一种新的基于模型的语音分析/合成系统”，载于*ICASSP '85. IEEE*国际声学、语音与信号处理会议，第10卷，（美国佛罗里达州坦帕），第513–516页，电气与电子工程师协会，1985年。[13] J. Engel、C. Resnick、A. Roberts、S. Dieleman、M. Norouzi、D. Eck 和 K. Simonyan，“使用WaveNet自编码器合成音符”，载于国际机器学习会议，第1068–1077页，PMLR，2017年。

[14] D. Griffin 和 Jae Lim，“从修改后的短时傅里叶变换进行信号估计”，*IEEE*《声学、语音与信号处理汇刊》，第32卷，第236–243页，1984年4月。[15] A. Natsiou 和 S. O' Leary，“一种用于梅尔频谱图逆变换的正弦信号重建方法”，载于*2021*年*IEEE*国际多媒体研讨会（*ISM*），（意大利那不勒斯），第245–248页，IEEE，2021年11月。[16] N. Srivastava、G. Hinton、A. Krizhevsky、I. Sutskever 和 R. Salakhutdinov，“随机失活：一种防止神经网络过拟合的简单方法”，《机器学习研究杂志》，第15卷，1958，2014年，第1929页–。[17] F. Girosi、M. Jones 和 T. Poggio，“正则化理论与神经网络架构”，《神经计算》，第7卷，第219–269页，1995年3月。[18] S. C. Wong、A. Gatt、V. Stamatescu 和 M. D. McDonnell，“理解用于分类的数据增强：何时进行变形？”，载于*2016*年国际数字图像计算：技术与应用会议（*DICTA*），（澳大利亚黄金海岸），第1–6页，IEEE，2016年11月。[19] G. Orr 和 K.-R. M¨uller，编，《神经网络：实用技巧》。《计算机科学讲义》第1524卷，柏林；纽约：Springer，1998年。[20] T. Ko、V. Peddinti、D. Povey 和 S. Khudanpur，“用于语音识别的音频增强”，载于*Interspeech*，第3586–3589页，ISCA，2015年。[21] S. S. Stevens、J. Volkmann 和 E. B. Newman，“心理物理音高量值测量标度”，《美国声学学会杂志》，第8卷，第3期，第185–190页，1937年。[22] D. P. Kingma 和 J. Ba，“Adam：一种随机优化方法”，2017年1月。编号：arXiv:1412.6980 arXiv:1412.6980 [cs]。[23] J. Davis 和 M. Goadrich，“精确率-召回率与ROC曲线之间的关系”，载于第*23*届国际机器学习会议论文集——*ICML '06*，（宾夕法尼亚州匹兹堡），第233–240页，ACM出版社，2006年。[24] A. van den Oord、S. Dieleman、H. Zen、K. Simonyan、O. Vinyals、A. Graves、N. Kalchbrenner、A. Senior 和 K. Kavukcuoglu，“WaveNet：一种原始音频生成模型”，载于第*9*届*ISCA*语音合成研讨会，第125–125页，2016年。[25] P. Esling、A. Chemla-Romeu-Santos 和 A. Bitton，“生成式音色空间：利用感知度量对变分自编码器进行正则化”，载于国际数字音频效果会议（*DAFx*）论文集，葡萄牙阿威罗。，2018年。[26] F. Roche、T. Hueber、M. Garnier、S. Limier 和 L. Girin，“让声音更具金属质感：利用变分自编码器实现对合成器声音音色的感知相关控制”，《国际音乐信息检索学会汇刊（*TISMIR*）》，第4卷，第52–66页，2021年。[27] G. Vilone 和 L. Longo，“根据输出格式对可解释人工智能方法进行分类”，《机器学习与知识提取》，第3卷，第3期，第615–661页，2021年。[28] G. Vilone 和 L. Longo，“可解释性概念与可解释人工智能的评估方法”，《信息融合》，第76卷，第89–106页，2021年。