基于改进高斯混合变分自编码器的半监督情感音乐生成

胥备, 刘桐

引用本文

胥备, 刘桐. 基于改进高斯混合变分自编码器的半监督情感音乐生成[J]. 计算机科学, 2024, 51(8): 281-296.

XU Bei, LIU Tong. Semi-supervised Emotional Music Generation Method Based on Improved Gaussian Mixture Variational Autoencoders [J]. Computer Science, 2024, 51(8): 281-296.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230500124

基于改进高斯混合变分自编码器的半监督情感音乐生成

胥 备[1,2] 刘 桐[1]
1 南京邮电大学计算机学院 南京 210023
2 江苏大数据安全与智能处理重点实验室 南京 210023
(xubei@njupt.edu.cn)

摘 要 音乐可以通过序列化的声音信息传递声音内容和情感。情感是音乐所表达的语义中的重要组成部分,因此,音乐生成技术不仅要考虑音乐的结构信息,还应融入情感元素。现有的情感音乐生成技术大多采用基于情感标注的完全监督方法,但音乐领域缺乏大量标准的情感标注数据集,且情感标签不足以表达音乐的情感特征。针对上述问题,提出了基于改进的高斯混合变分自编码器(Gaussian Mixture Variational Autoencoders,GMVAE)的半监督情感音乐生成方法(Semg-GMVAE),将音乐的节奏特征和调式特征与情感建立联系,同时向GMVAE中引入一种特征解纠缠机制来分别学习这两种特征的潜在变量表示,并对其进行半监督聚类推断。最后通过操纵音乐的特征表示,实现了针对快乐、紧张、悲伤、平静情感的音乐生成与情感转换。同时,针对GMVAE难以区分不同情感类别数据的问题,实验指出其关键原因是GMVAE证据下界中的方差正则项与互信息抑制项使得各类别的高斯分量分散性不足,从而影响学习表示的性能和生成的数据样本的情感质量。因此,Semg-GMVAE对这两项因子分别进行了惩罚和增强,并使用Transformer-XL作为编码器和解码器以提升在长序列音乐上的建模能力。基于真实数据集的实验结果表明,相比现有方法,Semg-GMVAE能够将不同情感的音乐在潜在空间中更好地分离,增强了音乐与情感的关联程度,并且能够有效对不同音乐特征进行解纠缠分离,最后通过改变特征表示更好地实现情感音乐生成或情感切换。

关键词:情感音乐生成;半监督生成模型;解纠缠表示学习;高斯混合变分自编码器;Transformer-XL
中图分类号 TP181

Semi-supervised Emotional Music Generation Method Based on Improved Gaussian Mixture Variational Autoencoders

XU Bei[1,2] and LIU Tong[1]
1 School of Computer Science,Nanjing University of Posts and Telecommunications,Nanjing 210023,China
2 Jiangsu Key Laboratory of Big Data Security & Intelligent Processing,Nanjing 210023,China

Abstract Music can transmit audio content and emotions through serialized audio features. Emotion is an important component in the semantic expression of music. Therefore,music generation technology should not only consider the structural information of music but also incorporate emotions. Most existing emotional music generation technologies use the complete supervised methods based on emotion labeling. However,the music field lacks a large number of standard emotional labeling datasets,and emotional labels are insufficient to express the emotional features of music. To solve these problems,this paper proposes a semi-supervised emotional music generation method(Semg-GMVAE) based on improved Gaussian mixture variational autoencoders(GMVAE),which connects the rhythm features and mode features of music with emotions,incorporates a feature disentanglement mechanism into GMVAE to learn the potential variable representations of these two features,and performs semi-supervised clustering inference on them. Finally,by manipulating the feature representation of music,our model can achieve music generation and emotion switching on happy,tense,sad,and calm emotions. Meanwhile,this paper conducts a series of experiments on the problem that GMVAE is difficult to distinguish different emotional categories of data. The key reason for the problem is that the variance regularization term and mutual information suppression term in the evidence lower bound of GMVAE make the Gaussian components of each category less dispersed,thus affecting the performance of learned representation and the quality of generation. Therefore,Semg-GMVAE penalizes and augments these two factors respectively,and uses Transformer-XL as the encoder and decoder to enhance the modeling capabilities on long sequence music. Experimental results based on real data show that,compared to existing methods,Semg-GMVAE achieves better separation of music with different emotions in potential space,enhances the correlation between music and emotions,effectively disentangles different music features,and finally achieves better emotional music genera-

到稿日期:2023-05-21 返修日期:2023-11-14
基金项目:江苏省高校自然科学基金面上项目(21KJB520017)
This work was supported by the Natural Science Foundation of the Jiangsu Higher Education Institutions of China(21KJB520017).
通信作者:刘桐(liutong9986@163.com)

tion and emotion switching by changing the feature representation.

Keywords Emotional music generation, Semi-supervised generative models, Disentangled representation learning, Gaussian mixture variational autoencoders, Transformer-XL

1 引言

音乐作为艺术领域中一种重要的表达方式，体现了一系列人类特有的思维模式。音乐与自然语言一样，可以通过序列化的信息传达内容和情感。但不同于自然语言，情感是音乐所传达信息的重要组成部分。音乐情感不仅与音乐内在结构密切相关，且受到随时间传播的音乐元素序列诱发。因此，音乐的自动生成不仅要符合乐理规则，更需要揉入情感元素以激发创造性。基于情感的音乐生成被广泛应用于电影或游戏的特定场景背景音乐生成、基于音乐的心理治疗等应用中。

近年来，自动音乐生成技术取得了重大发展，已有大量生成模型能够很好地学习音乐的结构特征，输出丰富的音乐样本。但较少有研究将情感元素作为音乐生成的条件。一种典型的面向情感的音乐生成方法是利用循环神经网络(RNN)[1-2]、Transformer[3-4]等序列对序列(Seq2Seq)模型同时对音乐序列和情感标签编码，学习特定情感的特征表示，然后将该特征表示解码生成相应情感的音乐，或对生成后的音乐以音乐情感分类器为辅助进行控制。为了得到更好的情感语义表达，部分研究利用变分自编码器(VAE)[5]学习情感标签嵌入并对数据的潜在空间进行建模[6]，将数据的分布映射到一个低维连续空间中，然后从潜在空间中采样，以生成多样化和连续的样本。但上述方法采用完全监督的训练并直接基于标签生成特定情感的音乐，存在两方面的缺陷。一方面，完全监督训练会过度依赖有标签数据，但音乐领域缺乏大量标准的情感数据集，且情感具有很强的主观性，在标注数据时可能产生较多的标签噪声。另一方面，情感标签不足以表达音乐的情感特征，直接基于标签生成音乐忽略了情感与音乐元素之间的特征关系，无法对生成的音乐解释其内在的情感表征。

因此，最近的研究 Music FaderNets[7]利用高斯混合变分自编码器(GMVAE)[8]通过半监督方法学习了音符特征和节奏特征在 Russell 二维情感空间[9]的 Arousal 维度上的潜在空间表示，并利用这两种特征实现了对 Arousal 维度的控制，从而生成特定情感的音乐。GMVAE 可以将 VAE 中的单一潜在分布扩展到多个高斯分量上，每个高斯分量代表一类数据。然而，GMVAE 存在一定的瓶颈，可能会出现模式坍塌问题[10]。一方面，当数据过于相似时，在训练过程中各个高斯分量通常无法很好地分离，难以区分不同类别数据，从而影响学习表示的性能和生成的数据样本的情感质量。另一方面，现有的基于 VAE 的相关音乐生成模型大多采用循环神经网络(RNN)作为编码器和解码器[5,7,11-12]。而音乐属于长序列数据，RNN 在此类数据上建模能力有限，容易丢失上下文依赖关系，并且存在梯度消失或梯度爆炸的问题。

针对上述问题，本文提出了一种基于类别分散型高斯混合变分自编码器的半监督情感音乐生成模型(Semg-GMVAE)，用于学习不同音乐特征对情感的影响，生成具有长期依赖结构和特定情感的音乐，并通过操控音乐特征来实现音乐情感的转换。本文的具体贡献如下：

(1)提出了一种具有方差惩罚和互信息增强的 GMVAE 半监督情感音乐生成模型，能够保证不同情感的音乐在潜在空间中取得更好的分离效果，增强了音乐与情感信息的相关程度，使得生成的音乐更贴近目标情感，并提高了半监督模型的鲁棒性和泛化能力。

(2)提出了一种具有独立编码器约束和生成对抗损失函数约束的特征解纠缠机制从音乐序列中分离学习关于节奏特征和调式特征的潜在变量表示，通过操控这两种特征表示可实现特定情感音乐的生成与转换。

(3)引入了 Transformer-XL 网络[13]作为 GMVAE 的编码器和解码器。基于其段级循环机制和相对位置编码，能够有效学习音乐序列中更长的上下文依赖结构，并且增强了模型对不同特征的关注能力，提高了模型的表现力。

2 相关工作

2.1 情感模型

心理学家们提出了一系列情感模型来描述情感类别。常用的情感模型分为两类：分类情感模型和维度情感模型。典型的分类情感模型如 Hevner 模型[14](它包含 8 个类别的 67 个情感形容词)以及日内瓦情感音乐量表[15](它包含 9 个类别的 45 个情感标签)。维度情感模型通过设置若干连续的情感维度以表达情感在某一维度上的连续变化，较多应用于回归任务中。例如 Thayer 模型[16]给出了两种情感维度：活跃度(Energetic)和紧张度(Tense)。PAD 模型[17]给出了 3 种情感维度：愉悦度(Pleasure)、唤醒度(Arousal)、优势度(Dominance)。Russell[9]的二维情感空间是音乐情感任务中的常用模型，它将情感的维度分为唤醒度(Arousal)和效价度(Valence)，如图 1 所示。其中 Arousal 表示用户神经的激活水平及兴奋程度(高兴奋为正值，低兴奋为负值)；Valence 表示用户情感状态的积极性(正值)和消极性(负值)。

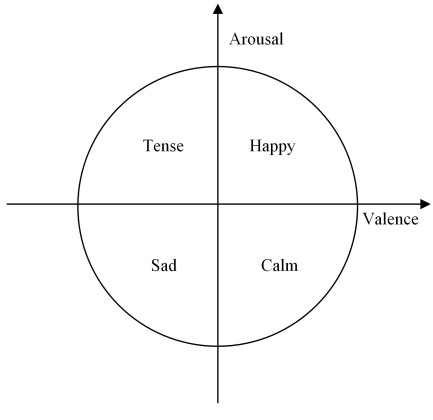


图 1 Russell 情感模型

Fig. 1 Russell emotion model

相比其他情感模型，Russell 模型具有直观、全面、可区分

度高的特点。因此，较多音乐数据集选择 Russell 模型进行情感标注。同时，音乐心理学等领域[18-19]利用 Russell 模型探究了音乐特征与情感之间的关系，对音乐的特征解纠缠任务起到了重要作用。因此，本文使用 Russell 情感模型表达音乐情感。为了更加具体地展现不同情感，我们对该模型的二维空间进行了离散化处理，得到了 4 种离散情感标签：快乐、紧张、悲伤、平静。

2.2　VAE 半监督生成模型

半监督生成模型一般可用于生成给定条件的样本，其中小部分有标签数据用于训练模型，而无标签数据用于辅助训练。相比完全监督的方法，半监督生成模型可以学习到更丰富的数据分布，适用性更广。半监督生成模型主要基于 VAE 和 GAN 框架实现[20]，目前已在图像分类、文本生成、语音合成等领域得到广泛应用。Kingma 等[21]最早提出了基于 VAE 的半监督生成模型(SSVAE)。该方法将类标签也视为一种服从类别分布的离散潜在变量，由编码器经过后验推断得到，与连续潜在变量共同生成数据，并通过有监督损失和无监督损失来训练模型。在最近的研究中，Habib 等[22]将具有注意力机制的序列对序列神经网络(Seq2Seq)与 SSVAE 相结合，实现了能够控制情感、基频和语速的语音生成任务。Cheung 等[23]将双向长短期记忆神经网络(BiLSTM)[24]与 SSVAE 相结合，利用乐谱上的音高和时间信息来生成小提琴指法。上述研究证明了 VAE 半监督生成模型的潜在空间能够提供更稳健的特征，仅需要少量监督即可达到控制的效果，相比完全监督的方法，能够达到更好的性能。但是，SSVAE 相当于以分类器的方式从输入数据中学习数据的类别信息，这会引入更多的学习参数，使得半监督模型过度依赖于有标签数据，并且 SSVAE 学习到的类别信息只局限于标签本身，无法代表数据关于某种类别的真正语义，在生成数据时也很难利用到这种微小的类别信息。针对以上问题，Li 等[25]提出了一种半监督解纠缠 VAE(SDVAE)，该模型将输入数据编码为解纠缠表示和不可解释表示。其中解纠缠表示用于捕获类别信息，不可解释表示用于重建数据，并利用类别信息将解纠缠表示正则化为等式约束，可以直接从解纠缠表示获得分类结果，从而减少分类器在半监督学习中的使用，但仍然忽略了类别信息在生成时的不可用性。因此，Joy 等[26]详细分析了将标签作为潜在变量生成数据的不足之处，并提出了一种专门用于捕获不同标签特征的 VAE 生成模型(CCVAE)，该模型将潜在变量设计为与标签相关的特征表示，然后使用分类器来推断潜在变量的标签值，从而影响潜在变量的结构，并且每种类型的标签都有独立的潜在表示，可以操纵特定标签的生成，但这也引入了更多的分类器来控制潜在变量的表达。另一种方法则是利用 GMVAE[8]来解决该问题。GMVAE 向 VAE 中引入了高斯混合模型(GMM)[27]，可将 VAE 中的单峰高斯分布扩展到具有多个分量的混合高斯上，使得潜在空间分离为多个不同类簇，因此可直接从潜在变量中推断数据的类别，而无需引入额外的神经网络学习类别信息，在半监督生成任务中取得了优秀的性能。在已往的研究中，Luo 等[28]利用 GMVAE 实现了控制音高和音色的音频音乐生成。Tan 等[7]提出的模型 Music FaderNets 也利用 GMVAE 实现了 MIDI 音乐音符和节奏的可控生成。

然而，GMVAE 还存在一定的模式坍塌问题，即代表各个类簇的高斯分量在训练过程中可能无法获得较高的分散性，这会影响不同类别数据学习表示的性能和生成的数据样本的情感质量。因此，本文提出了一种改进的 GMVAE 半监督情感音乐生成模型(Semg-GMVAE)，通过优化其目标函数使得模型更容易区分不同情感类别的高斯分量，提高生成音乐的情感意义。除此之外，还将 GMVAE 模型与 Transformer-XL[13]相结合以生成更具长期依赖结构的音乐内容。

2.3　解纠缠可控生成

在传统的生成模型中，数据通常表示为高维特征向量。这些特征可能存在高度相关性，使得模型难以捕捉数据中的关键因素或不同因素之间的关系，并且此类模型在生成过程中也无法进行控制，限制了模型的可解释性和可控性。解纠缠机制可以学习数据的独立因素，将数据解耦成多个部分，使得每个部分只表示数据的一个方面。例如在图像数据中，可以将图片中的颜色、纹理、形状、光照等多种因素分离出来，每种因素可表示为特征的一个维度，从而进行独立的控制和操作[29-30]。

VAE 是实现解纠缠可控生成的有效框架[31-33]。其先验高斯分布的各向同性，使得潜在变量不同维度之间相互独立，可以分离出数据中的不同变化因素，但关键在于向模型中纳入适当的归纳偏置[34]，如多任务损失函数约束[12,35]、编码器约束[28,36-37]、解码器约束[38]等。Yang 等[35]提出了一种显示约束条件变分自编码器 EC^2-VAE，通过多任务损失函数实现音高特征和节奏特征的解耦分离。具体地，该模型首先从编码器学习到的整体潜在变量中分离出音高变量和节奏变量，然后利用节奏解码器和全局解码器分别重构节奏特征和原始输入，并计算相应的重构损失进行约束。在其后续的研究中，Poly-Dis 模型[34]被提出用于对复调音乐的和弦与纹理解耦分离。此外，Music FaderNets[7]还在解纠缠机制中引入了约束编码器的归纳偏置，为音符特征和节奏特征单独设计了编码器用于学习相应的潜在变量表示。这种方式可防止单个编码器中各特征之间发生竞争而导致无法充分发挥解纠缠能力的问题。

在成功解纠缠学习出不同特征的潜在变量后，可通过改变潜在变量来控制样本的生成，具体有交换、采样、插值等方法[39]。交换即选择两个样本交换其部分表征以产生两个新的样本。EC^2-VAE 和 Poly-Dis 都采用这种方法实现了音乐的可控生成，即保持一种特征潜在变量不变，而另一种特征潜在变量使用其他音乐来代替，从而生成新的音乐。实验表明，生成的音乐依然具有未被改变的特征，而改变后的特征与另一首音乐一致，这也证明了解纠缠学习的成功。然而交换的控制策略依赖于现有的样本，无法展现生成样本的多样性，因此可以利用采样的方法从潜在空间中重新获取新的潜在变量。这种方法非常适用于具有明确类别划分的潜在空间。例如，Luo 等[28]提出的 GMVAE 音频音乐生成模型，通过解纠缠机制分别学习了不同类别音高和音色的潜在空间，在生成特定音高或音色的音乐时，从目标类别的潜在空间重新采样相应的潜在变量解码输出。采样的方法由于具有随机性，

容易产生不平滑的数据点。插值法则是基于 VAE 潜在空间的连续性和平滑性，沿着潜在空间中两点之间的路径进行插值来生成新的数据点，可以产生平滑的过渡效果[40]。其生成的新样本在语义上会更加连贯、自然，有助于更精确地控制潜在空间中的变化。该方法也在 EC²-VAE，MusicVAE[11] 和 Music FaderNets 中得到了广泛应用。

基于上述有效的解纠缠方法和控制策略，本文在 Semg-GMVAE 模型中定义了一种特征相关的潜在空间，通过引入编码器约束学习音乐关于节奏特征和调式特征的潜在变量表示，并采用具有生成对抗机制的多任务损失函数约束以保证每一种潜在变量只包含与其相对应特征的信息，去除其他无关特征信息，进一步增强解纠缠效果。此外，在完成特征解纠缠的基础上，通过插值法操控潜在空间实现特定情感音乐的生成和变换。

3 半监督情感音乐生成方法

本文提出的 Semg-GMVAE 模型基于 Transformer-XL 网络的高斯混合变分自编码器(GMVAE)，如图 2 所示，主要包含 3 个模块：

(1)编码器模块：接收以小节分段的音乐输入序列，利用 Transformer-XL 编码器学习潜在分布的均值和方差。

(2)解码器模块：从潜在分布中采样潜在变量表示，利用 Transformer-XL 解码器重构音乐序列。

(3)高斯混合模块：利用高斯混合模型(GMM)以半监督学习方法学习潜在变量的类别信息，完成音乐数据的情感聚类。

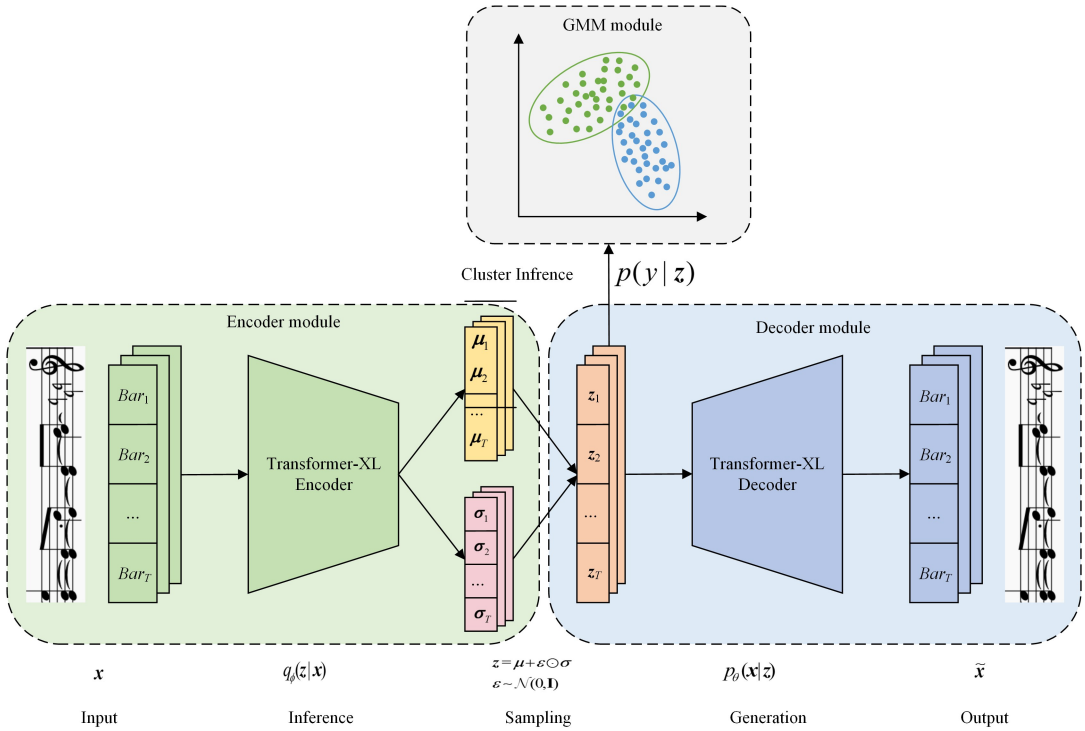


图 2 Semg-GMVAE 模型结构

Fig. 2 Structure of Semg-GMVAE model

3.1 半监督高斯混合变分自编码器

高斯混合变分自编码器(GMVAE)由变分自编码器(VAE)和高斯混合模型(GMM)构成。GMM 能够将 VAE 的连续潜在空间扩展为具有多个高斯分量的离散空间，每个高斯分量为一个类簇，包含了同一种类别的潜在变量表示。Semg-GMVAE 情感音乐生成模型旨在基于该特性以半监督方法对音乐特征的潜在变量表示进行情感类别的推断，将其投影到对应类别的高斯分量中。在生成阶段，只需从目标类别的高斯分量中采样潜在变量，然后解码生成相应情感的音乐。半监督 GMVAE 可以在不引入额外的神经网络分类器的情况下，利用少量有标签数据作为监督信号来指导模型学习各个类别的特征表示和边界，并且无标签数据提供了更广泛的数据分布，通过有标签数据的引导，可将相似样本归为同一类别。

在实现上，GMVAE 可分为推断网络和生成网络，其概率图模型如图 3 所示。对于推断网络，当给定音乐输入序列 $\boldsymbol{x}$，利用编码器学习任一特征的潜在变量表示 $\boldsymbol{z}$，并引入分类变量 y 来推断 $\boldsymbol{z}$ 的情感类别；对于生成网络，则可根据分类变量 y 从目标类别的高斯分量中采样潜在变量 $\boldsymbol{z}$，利用解码器基于该潜在变量重构输入数据 $\boldsymbol{x}$。其中，生成过程可按式(1)所示的联合概率分布定义为：

$$p_\theta(\boldsymbol{x},\boldsymbol{z},y)=p_\theta(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z}|y)p(y) \tag{1}$$

其中，$p(y)=\mathrm{Cat}(y|1/K)$，为具有 K 个类别的类别分布。$p(\boldsymbol{z}|y)=\mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_y,\boldsymbol{\sigma}_y^2)$，代表特定类别高斯分量的潜在分布，具有可学习的均值 $\boldsymbol{\mu}_y$ 和方差 $\boldsymbol{\sigma}_y^2$。$p_\theta(\boldsymbol{x}|\boldsymbol{z})$是以 θ 参数化的神经网络，用于解码生成样本数据。对于推断过程，GMVAE 和 VAE 一样通过变分推断 $q_\phi(\boldsymbol{z},y|\boldsymbol{x})$来近似真实后验分布 $p(\boldsymbol{z},y|\boldsymbol{x})$。根据平均场近似理论[10]，$q_\phi(\boldsymbol{z},y|\boldsymbol{x})$可进一步按式(2)因子分解为：

$$q_\phi(\boldsymbol{z},y|\boldsymbol{x})=q_\phi(\boldsymbol{z}|\boldsymbol{x})q_\phi(y|\boldsymbol{x}) \tag{2}$$

其中，$q_\phi(\boldsymbol{z}|\boldsymbol{x})=\mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu}_x,\boldsymbol{\sigma}_x^2)$，是通过 ϕ 参数化的编码器网络学习到的潜在分布，在训练过程中逐渐逼近相应类别下的高斯分布 $p(\boldsymbol{z}|y)$。类似地，$q_\phi(y|\boldsymbol{x})$用于从输入 $\boldsymbol{x}$ 中学习类别信息，但需要额外的神经网络来拟合，这会给模型引入更多的参数，并且更加依赖于有标签数据。因此，可根据 VaDE[41] 中的推导过程将 $q_\phi(y|\boldsymbol{x})$近似为 $p(y|\boldsymbol{z})$，近似方式如式(3)所示：

$$q_\phi(y|\boldsymbol{x})=\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[p(y|\boldsymbol{z})]\approx\frac{1}{N}\sum_{n=1}^{N}\frac{p(\boldsymbol{z}|y)p(y)}{\sum_{\hat{y}=1}^{K}p(\boldsymbol{z}|\hat{y})p(\hat{y})} \tag{3}$$

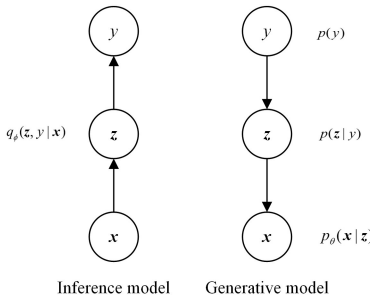


图 3　GMVAE 概率图模型

Fig. 3　Probabilistic graphical model of GMVAE

该方法利用了 GMM 的特性，并根据贝叶斯原理推断数据是由哪个高斯分量所代表的潜在分布生成的，从而有效划分数据类别，并且无需引入其他的神经网络。最终，式(3)通过计算 $p(y|\boldsymbol{z})$期望的蒙特卡洛估计来近似 $q_\phi(y|\boldsymbol{x})$，其中 N 为用于蒙特卡洛估计的样本数。在实现时，可利用已知类别的数据来初始化每个高斯分量的均值 $\boldsymbol{\mu}_y$ 和方差 $\boldsymbol{\sigma}_y^2$，然后根据这些初始化的参数来训练 GMVAE。在训练过程中，这些参数会不断被优化，直到能够准确地反映数据的类别信息。同时，未标记的数据也被用来更新高斯分量的参数，从而进一步提高对类别的学习能力。

与 VAE 一样，GMVAE 也通过最大化对数似然函数的证据下界(ELBO)来优化模型参数。根据上述的生成和推断过程，对数似然函数 $\log p(\boldsymbol{x})$可定义为：

$$\log\ p(\boldsymbol{x})\geqslant\mathbb{E}_{q_\phi(\boldsymbol{z},y|\boldsymbol{x})}\left[\log\frac{p_\theta(\boldsymbol{x},\boldsymbol{z},y)}{q_\phi(\boldsymbol{z},y|\boldsymbol{x})}\right]=\mathcal{L}_{\mathrm{ELBO}} \tag{4}$$

并且在使用无标签数据和有标签数据的半监督方法下，$\mathcal{L}_{\mathrm{ELBO}}$可进一步扩展为：

$$\begin{aligned}\mathcal{L}_{\mathrm{ELBO}}&=\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})]-\begin{cases}\mathcal{L}_{\mathrm{KL}}^{\mathrm{unsup}}, & \text{unsupervised}\\ \mathcal{L}_{\mathrm{KL}}^{\mathrm{sup}}, & \text{supervised}\end{cases}\\ \mathcal{L}_{\mathrm{KL}}^{\mathrm{unsup}}&=\mathbb{E}_{q_\phi(y|\boldsymbol{x})}[\mathrm{KL}[q_\phi(\boldsymbol{z}|\boldsymbol{x})\parallel p(\boldsymbol{z}|y)]]+\mathrm{KL}[q_\phi(y|\boldsymbol{x})\parallel p(y)]\\ \mathcal{L}_{\mathrm{KL}}^{\mathrm{sup}}&=\mathrm{KL}[q_\phi(\boldsymbol{z}|\boldsymbol{x})\parallel p(\boldsymbol{z}|y)]\end{aligned} \tag{5}$$

其中，$\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})]$表示生成真实数据的概率值，可通过最小化交叉熵重构损失函数实现；$\mathcal{L}_{\mathrm{KL}}^{\mathrm{unsup}}$为无监督学习下的 KL 散度损失，需要最小化编码器学习的潜在分布为与真实类别下的潜在分布之间的 KL 散度以及推断类别的后验分布与假设的均匀类别分布之间的 KL 散度；$\mathcal{L}_{\mathrm{KL}}^{\mathrm{sup}}$为有监督学习下的 KL 散度损失，只需最小化前者的 KL 散度即可。

3.2　方差项惩罚和互信息增强

在使用上述 GMVAE 模型对数据进行聚类推断时，由于数据的相似性过高以及高斯混合分量数量不足，可能无法学习到更复杂的分布。对应到潜在空间中，每个高斯分量在训练过程中难以分离，往往具有较为接近的均值和方差，甚至多个高斯可能坍塌为单一高斯。该问题被称为模式坍塌，是 GMVAE 较难区分不同类别数据的重要原因。

为了缓解 GMVAE 模式坍塌问题，本文对 3.1 节中的 ELBO 中做了进一步分析，认为其中的 KL 散度正则化项是引起该问题的主要原因。因此，对于 $\mathcal{L}_{\mathrm{KL}}^{\mathrm{unsup}}$中潜在分布 KL 散度损失项 $KL_z=\mathbb{E}_{q_\phi(y|\boldsymbol{x})}[KL[q_\phi(\boldsymbol{z}|\boldsymbol{x})\parallel p(\boldsymbol{z}|y)]]$，可重新分解为：

$$\begin{aligned}KL_z&=\sum_{\hat{y}=1}^{K}q_\phi(\hat{y}|\boldsymbol{x})\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\log\frac{q_\phi(\hat{y}|\boldsymbol{x})}{p(\boldsymbol{z}|\hat{y})}\\&=E_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\log\frac{q_\phi(\hat{y}|\boldsymbol{x})}{\prod_{\hat{y}=1}^{K}p(\boldsymbol{z}|\hat{y})^{q_\phi(\hat{y}|\boldsymbol{x})}}\end{aligned} \tag{6}$$

其中，$\mathbb{E}_{q_\phi(y|\boldsymbol{x})}=\sum_{\hat{y}=1}^{K}q_\phi(\hat{y}|\boldsymbol{x})=1$，$p(\boldsymbol{z}|\hat{y})$为类别 $\hat{y}$ 下的高斯分量，具有均值 $\boldsymbol{\mu}_{\hat{y}}$和方差 $\boldsymbol{\sigma}_{\hat{y}}^2$。并假设所有类别的高斯分量的方差都相等。因此，根据高斯分布概率密度函数，$p(\boldsymbol{z}|\hat{y})$可表示为：

$$p(\boldsymbol{z}|\hat{y})=\frac{1}{\sqrt{2\pi\boldsymbol{\sigma}_y^2}}\exp\left(-\frac{(\boldsymbol{z}-\boldsymbol{\mu}_{\hat{y}})^2}{2\boldsymbol{\sigma}_y^2}\right) \tag{7}$$

令 $\prod_{\hat{y}=1}^{K}p(\boldsymbol{z}|\hat{y})^{q_\phi(\hat{y}|\boldsymbol{x})}$为 $f(\boldsymbol{x},\hat{y},\boldsymbol{z})$，结合式(7)，可表示为：

$$\begin{aligned}f(\boldsymbol{x},\hat{y},\boldsymbol{z})&=\frac{1}{\sqrt{2\pi\boldsymbol{\sigma}_y^2}}\exp\left(-\frac{1}{2\boldsymbol{\sigma}_y^2}\sum_{\hat{y}=1}^{K}q_\phi(\hat{y}|\boldsymbol{x})(\boldsymbol{z}-\boldsymbol{\mu}_{\hat{y}})^2\right)\\&=\frac{1}{\sqrt{2\pi\boldsymbol{\sigma}_y^2}}\exp\left[-\frac{1}{2\boldsymbol{\sigma}_y^2}(\boldsymbol{z}-\mathbb{E}_{q_\phi(y|\boldsymbol{x})}\boldsymbol{\mu}_y)^2\right]\cdot\\&\quad\exp\left[-\frac{1}{2\sigma_y^2}(\mathbb{E}_{q_\phi(y|\boldsymbol{x})}\boldsymbol{\mu}_y^2-\mathbb{E}_{q_\phi(y|\boldsymbol{x})}^2\boldsymbol{\mu}_y)\right]\end{aligned} \tag{8}$$

其中，乘积的第一项因子可以被认定为所有类别的平均高斯分量 $\bar{p}(\boldsymbol{z}|y)$；第二项因子则代表所有高斯分量均值的方差，可衡量这些高斯分量的离散度。结合式(6)和式(8)，可得到 KL_z的变形，如式(9)所示：

$$KL_z=KL[q_\phi(\boldsymbol{z}|\boldsymbol{x})\parallel\bar{p}(\boldsymbol{z}|y)]+\frac{1}{2\boldsymbol{\sigma}_y^2}(\mathbb{E}_{q_\phi(y|\boldsymbol{x})}\boldsymbol{\mu}_y^2-\mathbb{E}_{q_\phi(y|\boldsymbol{x})}^2\boldsymbol{\mu}_y) \tag{9}$$

当在训练过程中最大化 GMVAE 的 ELBO 时，其中的 KL 散度损失会最小化，以促进学习到的潜在分布接近于真实的数据分布。根据该事实，KL_z 中的方差项因子也会最小化，使得各个高斯分量均值的方差越来越小，从而导致各个高斯分量在潜在空间中联系越来越紧密，无法正确区分不同类别，引起 GMVAE 模式坍塌的问题。为缓解这种收缩趋势，一种可靠的方法是对该方差项进行惩罚，调整其权重以减小对潜在空间的影响，从而增加各个高斯分量均值的距离，提高分散性。为此，本文在 KL_z的方差项前添加了用于惩罚的超参数 α，如式(10)所示：

$$KL_z=KL[q_\phi(\boldsymbol{z}|\boldsymbol{x})\parallel\bar{p}(\boldsymbol{z}|y)]+\frac{\alpha}{2\boldsymbol{\sigma}_y^2}(\mathbb{E}_{q_\phi(y|\boldsymbol{x})}\boldsymbol{\mu}_y^2-\mathbb{E}_{q_\phi(y|\boldsymbol{x})}^2\boldsymbol{\mu}_y) \tag{10}$$

其中，$\alpha\in[0,1]$，当 α 越小，方差项在整个 KL 散度损失中的权重就会减小，使得模型在训练过程中一定程度上忽略了对

各个高斯分量均值的方差的优化，从而提升潜在空间的分散性。在之后的消融实验中，本文证明了合适的 α 值能够在情感音乐生成任务产生良好的作用。

除了分离各个类别的高斯分离能够提高聚类的准确性，在输入数据与类别信息之间建立稳固的联系也是有必要的。但是，研究发现 VAE 的 ELBO 存在一种信息抑制项削弱了输入数据与潜在变量的关系[42]，这项瓶颈同样存在于 GM-VAE 的输入数据与类别信息中。在模型训练过程中，通常需要计算一个批次数据的平均 ELBO 来进行梯度更新。因此，结合式(10)，$\mathcal{L}_{\text{ELBO}}$ 可扩展为：

$$\mathcal{L}_{\text{ELBO}}=\mathbb{E}_x[\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})]-KL_z-KL[q_\phi(y|\boldsymbol{x})\parallel p(y)]] \tag{11}$$

根据互信息关于 KL 散度的定义[43]，输入数据 $\boldsymbol{x}$ 和类别信息 y 之间的互信息 $MI(y,\boldsymbol{x})$ 可表示为：

$$MI(y,\boldsymbol{x})=E_x[KL[q_\phi(y|\boldsymbol{x})\parallel p(y)]]-KL[q_\phi(y)\parallel p(y)] \tag{12}$$

因此，$\mathcal{L}_{\text{ELBO}}$ 可改写为：

$$\mathcal{L}_{\text{ELBO}}=\mathbb{E}_x[\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})]-KL_z]-MI(y,\boldsymbol{x})-KL[q_\phi(y)\parallel p(y)] \tag{13}$$

由此可见，在训练过程中，输入数据和类别信息之间的互信息也会被最小化，导致这两个随机变量的之间的相关依赖性减小，降低了聚类的准确性。因此，本文提出对 ELBO 和互信息进行联合优化，即：

$$\mathcal{L}_{\text{ELBO}}+MI(y,\boldsymbol{x})=\mathbb{E}_z[\mathbb{E}_{q(z|x)}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})]-KL_z]-KL[q_\phi(y)\parallel p(y)] \tag{14}$$

这种方法仍然可以通过 $KL[q_\phi(y)\parallel p(y)]$ 来优化类别分布的学习。其中 $q_\phi(y)=\mathbb{E}_x[q_\phi(y|\boldsymbol{x})]$，为 $q_\phi(y|\boldsymbol{x})$ 的离散边缘概率分布，可通过一个批次内的数据点平均得到。这使得所有数据点上的类别分布与先验分布的差异越来越小。

至此，结合方差项惩罚与互信息优化，可得到 GMVAE 的无监督 KL 散度损失，如式(15)所示：

$$\mathcal{L}_{KL}^{\text{unsup}}=KL[q_\phi(\boldsymbol{z}|\boldsymbol{x})\parallel\bar{p}(z|y)]+\frac{\alpha}{2\boldsymbol{\sigma}_y^2}\cdot(\mathbb{E}_{q_\phi(y|\boldsymbol{x})}\boldsymbol{\mu}_y^2-\mathbb{E}_{q_\phi(y|\boldsymbol{x})}^2\boldsymbol{\mu}_y)+KL[q_\phi(y)\parallel p(y)] \tag{15}$$

通过这两种增强方法，能够确保不同类别的数据在潜在空间中更加远离，而相同类别的数据在潜在空间内更加紧凑。

3.3 特征解纠缠学习

为了在生成音乐时更好地表达情感特征，本文将音乐的节奏特征和调式特征作为与 Russell 情感模型的 Arousal 维度和 Valence 维度建立联系的纽带。在音乐特征和情感的联系中，音乐心理学相关研究[18]证明了 Arousal 维度通常与节奏特征有关，Valence 维度与调式特征有关。

根据不同音乐特征与情感之间的联系，Semg-GMVAE 进一步引入了一种具有独立编码器约束和生成对抗损失函数约束的特征解纠缠机制来学习音乐在二维情感空间上的特征表达，如图 4 所示。首先，两个结构相同的编码器网络 E_r 和 E_k 用于从音乐原始序列中学习节奏特征 $\boldsymbol{f}_r$ 的潜在变量 $\boldsymbol{z}_r$ 和调式特征 $\boldsymbol{f}_k$ 的潜在变量 $\boldsymbol{z}_k$，并分别推断每种潜在变量的情感类别，即对应情感维度的高维或低维空间。然后，两个单独的局部解码器 D_r 和 D_k 用于对 $\boldsymbol{f}_r$ 和 $\boldsymbol{f}_k$ 进行预测以正确学习相应特征的潜在表示，训练目标是最小化真实特征与预测特征之间的误差。此外，为确保每种潜在变量只包含与其相对应特征的语义信息，本文根据生成对抗网络的思想对模型增加了对抗性训练，以 $\boldsymbol{z}_r$ 和 $\boldsymbol{z}_k$ 分别作为 D_k 和 D_r 的输入，目的是将错误的信息反馈给每个解码器以产生虚假的预测结果 $\bar{\boldsymbol{f}}_r$ 和 $\bar{\boldsymbol{f}}_k$，使得 D_k 和 D_r 在训练过程中能够相互对抗，只针对相对应的潜在变量输入做出正确的特征预测，从而去除每种潜在变量中无关的特征信息。因此，该过程的训练目标是最大化真实特征与预测特征之间的误差，并且模型的参数更新只对编码器有效，可以保留解码器在进行准确预测时的性能。最后，两种潜在变量被合并(即 $\boldsymbol{z}=Concat[\boldsymbol{z}_r,\boldsymbol{z}_k]$)输入全局解码器 D_{global} 中以生成具有相应情感特征的音乐。由此，特征解纠缠的损失可定义为：

$$\mathcal{L}_{\text{Dis}}=\mathbb{E}_{q_{\phi_r}(\boldsymbol{z}_r|\boldsymbol{x})q_{\phi_k}(\boldsymbol{z}_k|\boldsymbol{x})}[\log p_{\varphi_r}(\boldsymbol{f}_r|\boldsymbol{z}_r)p_{\varphi_k}(\boldsymbol{f}_k|\boldsymbol{z}_k)]+\mathbb{E}_{q_{\phi_r}(\boldsymbol{z}_r|\boldsymbol{x})q_{\phi_k}(\boldsymbol{z}_k|\boldsymbol{x})}[\log(1-p_{\varphi_r}(\boldsymbol{f}_r|\boldsymbol{z}_k))(1-p_{\varphi_k}(\boldsymbol{f}_k|\boldsymbol{z}_r))] \tag{16}$$

其中第一项代表正确预测特征时的损失，第二项代表对抗性损失，这两项损失都采用交叉熵函数实现。

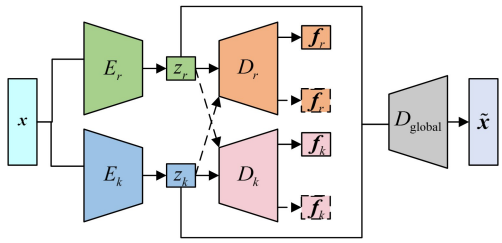


图 4 音乐特征解纠缠学习

Fig. 4 Music feature disentangled learning

综上所述，对于音乐特征 $\boldsymbol{f}_i$ 的潜在变量 $\boldsymbol{z}_i$ 和类别变量 y_i，半监督情感音乐生成模型 Semg-GMVAE 的目标函数可描述为：

$$\mathcal{L}_{\text{GMVAE}}=\mathbb{E}_{q_{\phi_r}(\boldsymbol{z}_r|\boldsymbol{x})q_{\phi_k}(\boldsymbol{z}_k|\boldsymbol{x})}[\log p_\theta(\boldsymbol{x}|\boldsymbol{z}_r,\boldsymbol{z}_k)]+\mathcal{L}_{\text{Dis}}-\sum_i\begin{cases}\mathcal{L}_{\text{KL}}^{\text{unsup},i}, & \text{unsupervised}\\ \mathcal{L}_{\text{KL}}^{\text{sup},i}, & \text{supervised}\end{cases} \tag{17}$$

3.4 编码器与解码器

以往的 VAE 模型及其变体常采用循环神经网络(RNN)构建编码器和解码器，但由于其在长序列数据上建模能力有限，也考虑将 Transformer[44] 与 VAE 模型相结合以生成音乐，如 Transformer-VAE[45]，MuseMorphose[46]。然而，Transformer 输入序列长度仍然受到内存的限制，每次都要基于整个序列进行自注意力计算。为了解决该问题，Semg-GMVAE 采用了图 5 所示的 Transformer-XL 的网络结构，利用其段级循环机制和相对位置编码可以学习更多音乐小节的上下文结构。

在该模型中，音乐序列按照小节被分为多个等长片段，即 $\boldsymbol{x}=[Bar_1,Bar_2,\cdots,Bar_T]$，其中 T 为划分的小节数。设 $\boldsymbol{h}_{\tau-1}^{n-1}\in\mathbb{R}^{L\times d}$ 为第 $\tau-1$ 个小节在第 $n-1$ 层自注意力的隐藏状态，其中 L 为每个小节的序列长度，d 为隐藏状态的维度。每当需要计算第 τ 个小节在第 n 层的隐藏状态 $\boldsymbol{h}_\tau^n$ 时，模型会考虑重用上一小节的隐藏状态，计算过程如式(18)所示：

$$\begin{aligned}&\widetilde{\boldsymbol{h}}_{\tau}^{n-1}=\text{Concat}[SG(\boldsymbol{h}_{\tau-1}^{n-1})\circ\boldsymbol{h}_{\tau}^{n-1}]\\&\boldsymbol{Q}_{\tau}^{n},\boldsymbol{K}_{\tau}^{n},\boldsymbol{V}_{\tau}^{n}=\boldsymbol{h}_{\tau}^{n-1}\boldsymbol{W}_{Q}^{\mathrm{T}},\widetilde{\boldsymbol{h}}_{\tau}^{n-1}\boldsymbol{W}_{K}^{\mathrm{T}},\widetilde{\boldsymbol{h}}_{\tau}^{n-1}\boldsymbol{W}_{V}^{\mathrm{T}}\\&\boldsymbol{h}_{\tau}^{n}=\text{Transformer-Layer}(\boldsymbol{Q}_{\tau}^{n},\boldsymbol{K}_{\tau}^{n},\boldsymbol{V}_{\tau}^{n})\end{aligned}\tag{18}$$

其中，$SG(\cdot)$表示 $\boldsymbol{h}_{\tau-1}^{n-1}$的梯度不会在计算下一小节时进行更新，$Concat[\cdot]$表示将两个隐藏状态矩阵按照小节长度拼接，$\boldsymbol{W}_{Q,K,V}$为可学习的模型参数。与 Transformer 不同的是，在进行注意力分数计算时，键值矩阵 $\boldsymbol{K}_{\tau}^{n}$ 和 $\boldsymbol{V}_{\tau}^{n}$ 取决于第 $n-1$ 层前一小节的隐藏状态 $\boldsymbol{h}_{\tau-1}^{n-1}$（如图 5 中神经网络结构的蓝线部分）与该层当前小节的隐藏状态 $\boldsymbol{h}_{\tau}^{n-1}$（如图 5 中神经网络的灰线部分）。通过这种方式，内存中只需缓存上一小节的隐藏状态，各个小节的隐藏状态可以在段与段之间建立循环连接，从而将上下文信息扩展至所有小节。

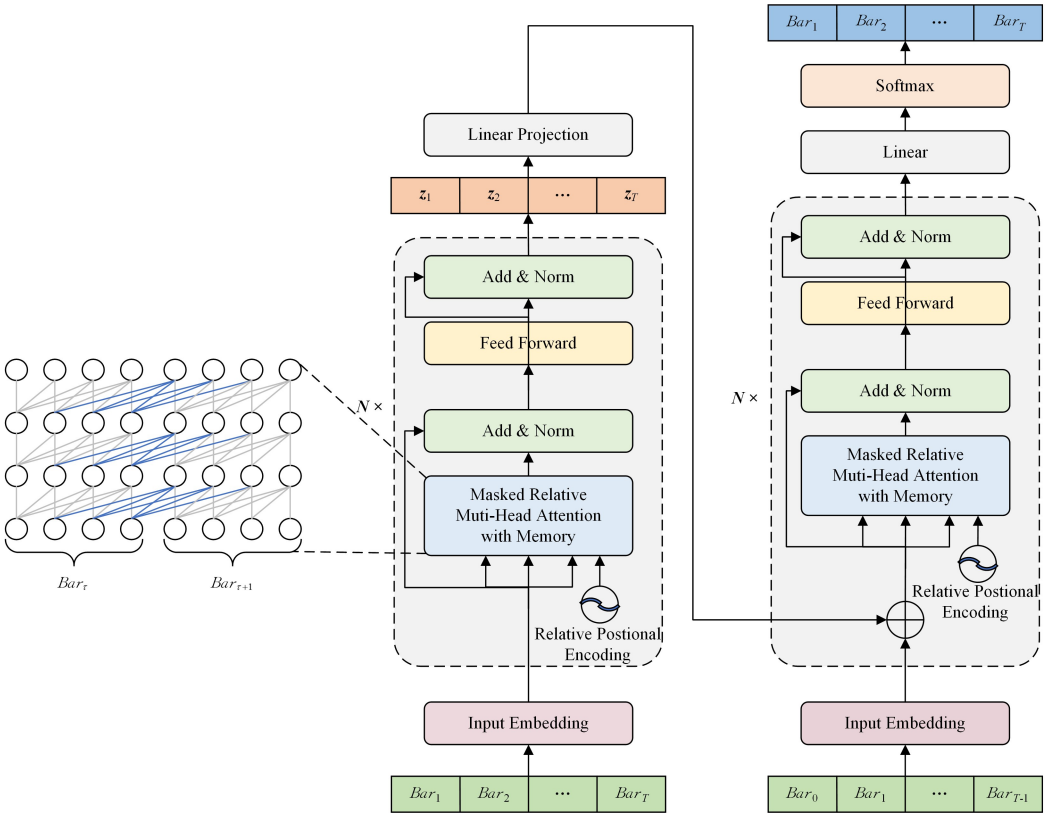


图 5　基于 Transformer-XL 的 Semg-GMVAE 模型结构

Fig. 5　Structure of Semg-GMVAE model based on Transformer-XL

Transformer 使用了一种位置编码来学习输入序列中的顺序和位置关系。然而，这种位置编码方式是绝对的，不同小节的同一位置的编码信息是相同的，会导致模型无法正确区分。因此，Transformer-XL 根据相对位置编码的思想，引入了一个无参数学习的正弦编码器矩阵 $\boldsymbol{R}\in\mathbb{R}^{L\times d}$，每一行代表两个位置的相对距离编码。在计算隐藏状态时，可动态地将相对距离注入注意力分数中。

将 Transformer-XL 与 GMVAE 模型结合时，需要从编码器输出隐藏状态中学习数据的潜在变量表示。根据 GMVAE 关于潜在分布 $q(\boldsymbol{z}|\boldsymbol{x})$的定义，可首先学习该分布均值和方差，然后从中采样得到潜在变量，其过程可表示为：

$$\begin{aligned}&\boldsymbol{h}_{\tau}^{\text{pool}}=\text{Avgpool}([\boldsymbol{h}_{\tau,1}^{N},\boldsymbol{h}_{\tau,2}^{N},\cdots,\boldsymbol{h}_{\tau,L}^{N}])\\&\boldsymbol{\mu}=\boldsymbol{h}^{\text{pool}}\boldsymbol{W}_{\mu}\log\boldsymbol{\sigma}^{2}=\boldsymbol{h}^{\text{pool}}\boldsymbol{W}_{\sigma}\\&\boldsymbol{z}=\boldsymbol{\mu}+\boldsymbol{\varepsilon}\odot\boldsymbol{\sigma}\quad\boldsymbol{\varepsilon}\sim\mathcal{N}(0,\boldsymbol{I})\end{aligned}\tag{19}$$

其中，$\boldsymbol{h}_{\tau}^{\text{pool}}\in\mathbb{R}^{d}$，代表第 τ 个小节整体隐藏状态，通过对所有位置的隐藏状态进行平均池化得到。$\boldsymbol{\mu},\boldsymbol{\sigma}\in\mathbb{R}^{T\times l}$，是从隐藏状态中学习到的关于所有小节潜在分布的均值向量和标准差向量，l 为设定的潜在变量维度，$\boldsymbol{W}_{\mu},\boldsymbol{W}_{\sigma}\in\mathbb{R}^{d\times l}$为参数矩阵。$\boldsymbol{z}\in\mathbb{R}^{T\times l}$则是根据重参数化技巧采样得到的关于所有小节的潜在变量。在进行解码生成时，潜在变量 $\boldsymbol{z}$ 通过线性层变换为模型统一维度，并与前一层输出隐藏状态相加后输入后续自注意力层中，从而利用有效的潜在变量信息生成最终音乐序列。

4　实验结果与分析

4.1　MIDI 音乐数据集

本文使用了公开的符号化 MIDI 音乐数据集 Lakh Pianoroll Dataset(LPD)[47]，该数据集包含 174 154 首多音轨 MIDI 音乐的钢琴卷帘矩阵。我们选取了其中的一个版本 LPD-5-full，并去除了重复内容、不包含钢琴音轨、总时长不到 1 min 的音乐，共得到 91 969 条数据，并按照 8∶1∶1 的比例划分为训练集、验证集和测试集。此外，该数据集能够与 Million Song Dataset(MSD)[48] 音频数据集中的音乐相匹配，通过 MSD 提供的元数据信息以及 Spotify 音乐社区提供的 API 服务[1)] 可获得 23 967 条具有情感标签的音乐数据。这些数据在 Russell 情感空间中具有标注好的 Arousal 值和 Valence 值，因此

1) http://developer.spotify.com

可将其量化到两个情感维度的高维或低维上，使用标签 A0，A1，V0，V1 进行标记，其中 0 代表低维，1 代表高维。将这 4 种标签两两组合可得到相对应的情感类别，即快乐、紧张、悲伤、平静。

4.2 音乐数据表示

为了使用 Semg-GMVAE 模型对 MIDI 音乐进行建模，需要一种合理的数据表示用作模型的输入。本文使用 Python 工具包 Pretty_midi 和 music21 来解析 MIDI 音乐中的结构信息，并根据 REMI[49] 将音乐编码为一系列 Token 事件的序列。每种事件有不同的取值，代表音符在时间顺序上发生的一次变化。例如 Bar 事件作为每个音乐小节的开始，Beat 事件作为每个节拍的开始，在每个 Beat 事件中添加相应的音符速率事件 Tempo 与音符事件 Pitch，Velocity，Duration。表 1 列出了每种事件的具体含义。对于 LPD 数据集，每首音乐最多包含 5 条音轨。Pitch，Velocity，Duration 是每条音轨独有的事件，因此可使用 Pitch-[Track]，Velocity-[Track]，Duration-[Track]表示在特定音轨 Track 上播放的音符事件。此外，我们将 EOS 标记添加到每条序列的末尾以表示序列的结束，将 PAD 标记用于序列的填充，最终产生的 Token 数量为 923。在加载数据时，将随机选取的连续音乐小节的数量 T 设置为 20，每个小节的最大序列长度 L 设置为 128。

表 1 MIDI 音乐事件信息

Table 1 MIDI music event information

事件名称	事件含义	Token 数量
Bar	表示一个音乐小节的开始	1
Beat	表示一个音乐小节中某个子节拍的开始，一个小节划分为 16 个 Beat	16
Tempo	表示音乐在某个时刻的节奏速率，范围设为 32～224 bpm，步长为 3 bpm	64
Pitch	表示一个音符的音高，范围设为 21～108	88 * 5
Velocity	表示一个音符的力度，范围设为 1～127，步长为 2	64 * 5
Duration	表示一个音符的持续时长，统一量化为占 n 个 Beat 的时长，$1 \leqslant n \leqslant 16$	16 * 5
EOS	表示序列的结束	1
PAD	填充标记	1

在特征解纠缠学习中，本文选择节奏特征和调式特征作为与情感建立联系的音乐特征。其中节奏特征 rhythm 使用长度为 $16 \times T$ 的一维向量表示，定义为音乐在每个节拍内同时播放的音符的数量。一段时间内音符出现的频率越高，说明节奏的激烈程度越高、表现更加急促；相反，频率越低，说明节奏更加温和平缓。这可用于衡量 Arousal 维度的情感。调式特征 key 使用长度为 T 的一维向量表示，定义为音乐在每个小节上可能出现的调式。调式分为 12 种大调（如 C 大调）和 12 种小调（如 c 小调），通常大调音乐拥有正向情感，而小调音乐拥有负向情感，可用于衡量 Valence 维度的情感。通过这两种特征在 Semg-GMVAE 中学习不同的潜在表示，可实现对音乐情感的把控。

4.3 实验环境与参数设置

本文使用 Pytorch 深度学习框架构建了 Semg-GMVAE 模型，并在配置了 4 个 NVIDIA Tesla P40 GPU 的服务器上进行了训练。对于 Semg-GMVAE 的网络结构，本文将 Transformer-XL 编码器与解码器的层数设为 12。每层自注意力机制的头数设为 8，隐藏层和潜在变量的维度分别设为 512 和 128。为了在 Arousal 和 Valence 情感维度上训练高斯混合模型，我们将每个维度的混合高斯数量设为 2，代表高维和低维，使用 Xavier 初始化[50]方法来初始化每个高斯分量的均值向量，初始标准差向量都设为恒定的 e^{-2}。在训练期间，将批量大小设为 16，使用 Adam 优化器来更新参数，并通过动态调整学习率促进模型收敛。在前 10 000 次迭代训练中，我们使用线性预热将学习率从 0 提高到 1×10^{-4}，在之后的 220 000 次迭代训练中，使用余弦退火衰减策略将学习率逐渐下降到 5×10^{-6}，经过 50 个训练周期后模型最终达到收敛。

4.4 实验分析

4.4.1 半监督生成效果分析

本小节实验评估了 Semg-GMVAE 模型的半监督生成效果，选取了 3 种采用完全监督方法的情感音乐生成模型以及 3 种半监督的情感音乐生成模型进行了性能对比。其中监督模型包括基于 Transformer 的方法 EMOPIA[4]、采用音乐特征与情感建立联系的 GRU 模型 EmotionBox[51] 和以情感标签作为条件的条件变分自编码器（CVAE）生成模型[6]。半监督模型采用了 2.1 节中提及的 SSVAE[21]，CCVAE[26]，FaderNets[7]。为了公平对比，所有半监督模型的主体网络结构采用 Transformer-XL 实现，并且除了 SSVAE 以外，都引入了特征解纠缠机制。此外，本文在 LPD 有标签数据上训练了 EMOPIA 提供的音乐情感分类模型，使用 8 个小节的音乐序列作为输入，经过 50 个训练周期，分类精度不低于对比模型。进一步地，利用训练好的分类器对所有模型生成的音乐分别在 Arousal 和 Valence 情感维度上进行了预测，通过准确率（Acc）、精确率（Pre）、召回率（Recall）、F1 值（F1）进行评估。

首先比较了各个生成模型在完全监督方法和半监督方法下对于目标情感音乐的生成效果。对比结果如表 2 和表 3 所列，分别对应 Arousal 和 Valence 情感维度。从表 2 和表 3 中可以发现，即使所有模型都采用了监督的训练方法，EMOPIA，EmotionBox 和 CVAE 的表现仍然不如其他生成模型。一方面是因为 GRU 和 Transformer 在学习音乐的上下文结构方面不如 Transformer-XL；另一方面是因为部分模型直接使用情感标签来表达音乐的情感，和其他特征解纠缠方法相比，缺乏一定的表示能力。实验结果还表明，在加入更多的无标签数据后，通过半监督方法训练得到的模型在所有分类指标上都远远优于完全监督的方法。这说明 VAE 生成模型及其变体能够有效地利用无标签数据学习更广泛的分布信息，并利用标签数据约束类别信息。总体上看，Semg-GMVAE 和 FaderNets 的生成效果比其他两种半监督模型更好。这是由于 GMVAE 模型能够引入标签信息，利用高斯混合分布将潜在空间分离为多个离散空间，有效地学习不同类别的数据分布。相比之下，其他的半监督 VAE 模型使用额外的神经网络来学习数据的类别信息，这可能会过度依赖有标签数据，使得分类器受到限制而无法充分地建模数据分布。此外，由于 Semg-GMVAE 在训练目标中对方差项和互信息项进行了增强，因此与 FaderNets 相比，会产生更高的分类精度。由此可以发现，Semg-GMVAE 能够更好地学习每个高斯分量上潜在变量表示，并准确地区分不同情感类别，该部分效果也会在后续小节中进行详细评估。

表 2　Arousal 情感维度上监督方法与半监督方法的性能对比

Table 2　Performance comparison of supervised and semi-supervised methods on Arousal emotion dimension

(%)

Models	Supervised				Semi-Supervised			
	Acc	*Pre*	*Rec*	*F*1	*Acc*	*Pre*	*Rec*	*F*1
EMOPIA	61.55	58.69	61.52	60.07	—	—	—	—
EmotionBox	60.06	57.35	58.71	58.02	—	—	—	—
CVAE	58.63	55.99	56.11	56.05	—	—	—	—
SSVAE	61.28	58.84	58.73	58.78	73.66	71.94	72.11	72.02
CCVAE	63.74	61.49	61.21	61.35	76.24	75.58	73.07	74.30
FaderNets	68.37	65.21	70.17	67.60	78.65	76.54	78.73	77.62
Semg-GMVAE	**70.52**	**66.92**	**73.77**	**70.18**	**82.54**	**79.56**	**84.60**	**82.00**

表 3　Valence 情感维度上监督方法与半监督方法的性能对比

Table 3　Performance comparison of supervised and semi-supervised methods on Valence emotion dimension

(%)

Models	Supervised				Semi-Supervised			
	Acc	*Pre*	*Rec*	*F*1	*Acc*	*Pre*	*Rec*	*F*1
EMOPIA	57.84	59.81	57.82	58.80	—	—	—	—
EmotionBox	57.18	59.06	57.66	58.35	—	—	—	—
CVAE	55.27	59.96	57.34	57.15	—	—	—	—
SSVAE	58.12	59.96	59.70	59.32	67.15	68.00	69.61	68.80
CCVAE	60.03	61.67	63.43	62.28	69.30	72.88	65.28	68.87
FaderNets	63.45	64.94	64.64	64.79	71.63	72.44	73.38	72.91
Semg-GMVAE	**66.63**	**68.93**	**65.28**	**67.05**	**75.87**	**77.37**	**75.78**	**76.57**

为了进一步展现半监督学习的良好性能，我们在不同数量的标签数据下测试了各个模型生成音乐的准确性。图 6 和图 7 分别描述了不同监督率 ρ 下音乐在 Arousal 和 Valence 情感维度上的准确率。其中 $\rho=M/(M+N)$，M 为使用的有标签数据的数量，N 为全部的无标签数据，因此 $0<\rho\leqslant 0.26$。从图中可以看出，Semg-GMVAE 和 FaderNets 在监督率为 0.05 的情况下就能够达到更高的准确率。这说明 GMVAE 仅利用少量有标签数据即可学习广泛的无标签数据在各个情感类别上的分布。此外，由于 Semg-GMVAE 在方差项和互信息方面得到了优化，在所有监督条件下其准确率与 FaderNets 相比都有很大的优势。这证明了在受到高斯分量之间缺乏分散性以及数据与情感类别联系不足的影响下，数据的类别推断会存在误差，进而降低生成音的情感质量。对于 SSVAE 和 CCVAE，结果显示它们在所有监督率上的准确率都远低于基于 GMVAE 的生成模型，这是因为 SSVAE 和 CCVAE 通过分类器学习关于某种情感有意义的潜在表示的能力有限。虽然从整体上看，SSVAE 和 CCVAE 的准确率提升幅度略大于 GMVAE 生成模型，并且仍有较大的上升空间，但这也从侧面说明了这两种模型过于依赖有标签数据释放的监督信号。在整个变化趋势上，CCVAE 的生成效果要优于 SSVAE。这是由于在实现 CCVAE 时引入了解纠缠学习，分类器能够从节奏特征和调式特征所代表的潜在变量中学习对应的情感类别，使得这两种潜在变量包含了所需要的情感信息。SSVAE 则是直接从原始输入中学习情感的类别信息，然后与潜在变量共同生成相应情感的音乐。在解码器生成过程中，这种微妙的情感信息通常容易被忽视。

综上所述，所有模型生成的音乐在 Arousal 维度上比在 Valence 维度上更容易区分。这是由于节奏特征更具普遍性和直接性，能够从原始输入数据中学习到良好的潜在表示。调式特征则需要经过复杂的乐理推断得到，因此模型缺乏这方面的能力。实验也表明了节奏特征与 Arousal 维度的相关性较高，更容易影响音乐的活跃程度。Valence 维度涉及的音乐情感极性，不能仅仅通过调式特征来呈现。这一结果也与 EmotionBox 中得出的结论一致。

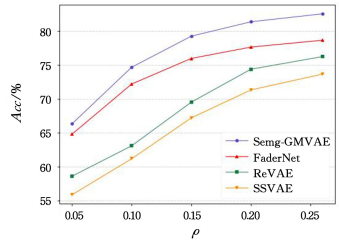


图 6　生成的音乐在 Arousal 维度上的准确率

Fig. 6　Accuracy of the generated music on Arousal dimension

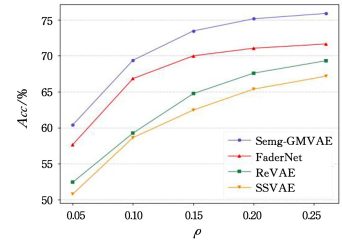


图 7　生成的音乐在 Valence 维度上的准确率

Fig. 7　Accuracy of the generated music on Valence dimension

4.4.2　方差惩罚和互信息增强性能分析

本小节主要阐述了 GMVAE 证据下界中的方差项和互信息抑制项如何对潜在空间产生影响，并且证明了在对这两种偏差进行合适的弥补后，模型能够对音乐数据进行合理的情感类别推断与生成。首先，我们比较了在不同权重参数（即 α 值）的惩罚力度下，各个高斯分量方差的变化趋势以及对生成音乐在情感准确率上的影响程度。实验结果如图 8 和图 9 中蓝线所示。为了能够更清楚地观察到高斯分量之间的变化，本文通过 T-SNE 降维算法[52]将高维潜在变量映射到二维空间中以进行可视化。实验结果如图 10 所示。

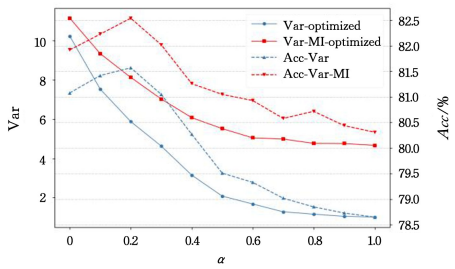


图 8 方差惩罚和互信息增强对 Arousal 维度的影响(电子版为彩图)

Fig. 8 Effect of variance penalty and mutual information enhancement on Arousal dimension

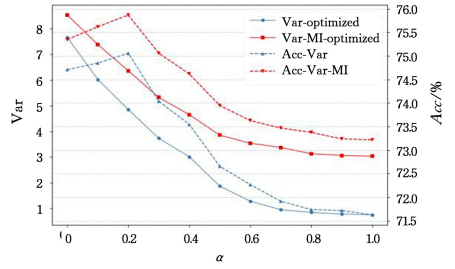


图 9 方差惩罚和互信息增强对 Valence 维度的影响(电子版为彩图)

Fig. 9 Effect of variance penalty and mutual information enhancement on Valence dimension

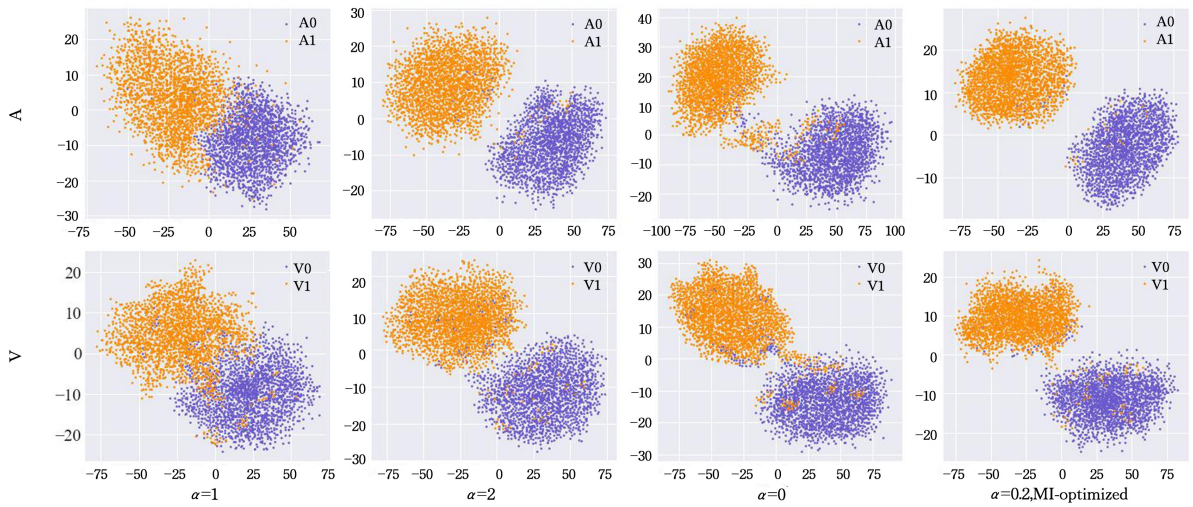


图 10 在方差惩罚和互信息增强影响下潜在空间的可视化效果

Fig. 10 Visualization of latent space under the influence of variance penalty and mutual information enhancement

正如 3.2 节所述,当 α 值越小,方差项对目标函数的整体影响就越小,因此能够产生较大的方差和较高的分类准确率,对应到潜在空间可视化图中,高斯分量的距离也会变大。这也对模型生成目标情感的音乐产生了良好的效果。但是,当 $\alpha \leqslant 0.2$ 时,虽然各个高斯分量的距离达到了最大,而生成音乐的情感准确率却下降了。从相应的潜在空间可视化图可以看出,两个高斯分量之间出现了不必要的聚类中心,并且有较多错误的类别推断,即同一类的数据被分配到了另一个类簇中。这是由于过高的惩罚力度可能会引起过拟合问题,模型将无法捕捉到数据的整体分布情况,导致信息丢失。因此根据实验结果,将 α 取值设定为 0.2,以发挥模型的最佳性能。此外,图 8 和图 9 中红色折线图进一步展示了在惩罚方差项基础上增强互信息后的性能变化。可以看出,在相同的权重控制下,优化互信息能够带来更大的方差和更高的准确率,并且在较小的 α 值下,互信息起到了更大的增强作用,弥补了方差惩罚力度不足的缺陷。从图 10 的可视化潜在空间中可发现,增强互信息后,不同类簇之间距离更远,相同类簇更加紧凑。

4.4.3 特征解纠缠性能分析

在实现基于情感生成音乐时,我们向 Semg-GMVAE 模型中引入了解纠缠机制,分别学习了节奏特征和调式特征在 Arousal 和 Valence 维度上的情感表示。本小节实验主要评估了改变特征表示对生成音乐的影响程度。理想情况下,模型学习到的两种特征表示都应该具有独立性。例如,当改变节奏的特征表示时,生成后的音乐的节奏模式也应该具有相应变化,而调式能够保持和原来相同。为了证明解纠缠学习的有效性,我们构建了以下几种解纠缠音乐生成模型作为对比模型,并在相同的实验环境设置下进行了对比实验。

(1)EC^2-VAE[35]:2.2 节中提及的音乐解纠缠生成模型。该模型采用基于双向 GRU 的 VAE 实现,从单个编码器学习到的整体潜在变量中分离了音高表示和节奏表示,然后利用节奏解码器和全局解码器分别重构节奏特征和原始输入。

(2)GAN-CVAE[53]:该模型定义了一个潜在表示独立于特征值的潜在空间,其思想是通过对抗机制学习一个不含有任何特征信息的广义潜在变量表示,而每个可控特征都划分为了不同的类别,会以类别标签的方式作为条件和潜在变量共同生成特定特征的音乐,并且同样由 GRU 进行构建。

(3)MuseMorphose[46]:该模型采用与 GAN-CVAE 类似的思想,但通过 Transformer 构建了主体网络,使得模型更注重整体的依赖信息,并且该模型研究了潜在变量与特征信息注入解码器中的不同方式,从而有效利用了已知的条件信息。

(4)Vanilla Semg-GMVAE:本文模型的简化版本,参考了文献[28]中的设计,除了 GMVAE 的证据下界,未使用任何损失函数进行约束。

(5)Semg-LSTM-GMVAE：本文模型的变体，使用双向长短期记忆神经网络(BiLSTM)构建了 GMVAE 模型。

基于以上对比模型，本文对特征的可控性进行了实验评估，所采用的实验方法是将测试集中的样本两两组成一对输入模型的编码器中，每一组的两个样本可分别记为 A 和 B。然后，在保持其中一种特征不变的情况下，互相交换另一种特征的潜在变量表示，最后比较解码器生成的音乐与原始样本之间特征的相似性。对于相似性的评估，我们计算了交换后的生成样本与 A 和 B 之间的余弦相似度，相似度的定义如式(20)所示：

$$sim(\boldsymbol{a},\boldsymbol{b})=\langle \boldsymbol{a},\boldsymbol{b}\rangle / \| \boldsymbol{a} \| \| \boldsymbol{b} \| \tag{20}$$

其中，$\boldsymbol{a}$ 和 $\boldsymbol{b}$ 分别代表生成样本和原始样本的特征向量，$\langle \cdot,\cdot \rangle$ 代表点积，$\| \cdot \|$ 代表特征向量的范数。

图 11 和图 12 的柱状图分别显示了在交换节奏特征和调式特征后，生成的样本与原始样本 A 和 B 之间特征相似度的变化。从图中可以看出，Semg-LSTM-GMVAE 和 Semg-GMVAE 模型经过合理的特征解纠缠分离，随意改变某一特征的潜在变量，并不会对另一种特征产生较大影响，并且当使用样本 B 的潜在变量替换样本 A 时，生成的音乐在节奏特征和调式特征方面都会模仿样本 B 的模式，产生了较高的特征相似度值。然而，当缺乏适当的特征约束时，例如 Vanilla Semg-GMVAE 和 EC^2-VAE 模型，虽然样本 A 改变后的特征与样本 B 之间产生了较高的相似性，但是其中一种特征的变化对另一种特征造成了一定程度的影响，导致其与原始样本 A 之间的特征相似度降低。这说明了节奏特征表示和调式特征表示未得到较好的解纠缠效果。

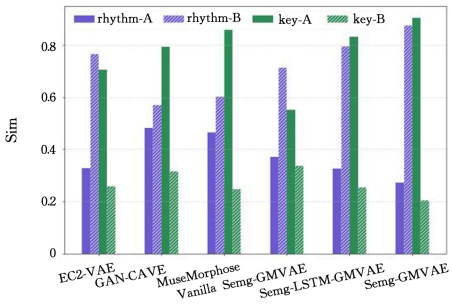


图 11　交换节奏特征表示后生成样本与原始样本之间的特征相似度

Fig. 11　Feature similarity between the generated sample and original sample after swapping rhythm feature representation

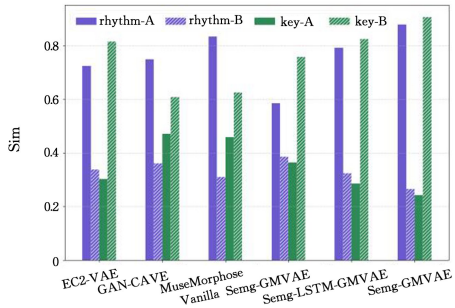


图 12　交换调式特征表示后生成样本与原始样本之间的特征相似度

Fig. 12　Feature similarity between the generated sample and original sample after swapping key feature representation

从图中还可以发现，对于 GAN-CVAE 和 Muse Morphose 模型，特征的控制产生的变化效果并不明显，相较于其他模型，GAN-CVAE 和 Muse Morphose 两种模型生成的样本之间的特征相似度仍然较高，而与替换样本 B 之间的特征相似度较低。这说明了定义特征不相关的潜在空间，将音乐的特征划分为多个类别区间，仅通过广义上的潜在变量表示和代表某种类别特征的标签信息来控制音乐的生成并不能够产生良好的效果，模型只会学习到一般的数据分布，并且可能会忽略这种微妙的特征信息。相比之下，Semg-GMVAE 的潜在空间直接反映了与特征相关的语义信息，使用表示特定特征的潜在变量可以提供更多的特征信息，使生成模型能够更精确地控制生成的数据。综上所述，对于一个通过特征控制的情感音乐生成模型，不仅需要良好的特征表示，并且各个特征表示应该相互独立，互不影响。

为了更清楚地看到改变特征表示后音乐的变化，我们从测试集中随机选取了两首音乐 A 和 B 输入最佳模型 Semg-GMVAE 中进行另一角度的控制生成评估。控制的方式仍然是交换音乐的特征表示，但比较了生成样本的钢琴卷帘矩阵和音高直方图。其中钢琴卷帘矩阵是一个二维矩阵。横轴代表以节拍为单位的时间步长，纵轴代表某一时刻演奏的所有音符音高，能够从整体上反映音符的变化。我们使用开源工具 pypianoroll[54] 提取了主旋律音轨上 8 个小节的钢琴卷帘矩阵，并对其进行了可视化操作。音高直方图是一种代表调式特征的数据结构，能够从整体上反映音乐所采用的调式。横轴代表 12 种不同的音阶，纵轴代表每一种音阶出现的频率。音高直方图可通过开源工具 Pretty_midi 计算得到。图 13 和图 14 分别展示了原始音乐样本 A 和 B 的钢琴卷帘矩阵和音高直方图。从图中可以发现，A 样本中每个小节中的音符密度较高，节奏比较急促，整体调式为 G 小调，代表了情感为 A1-V0 的音乐。B 样本中每个小节中的音符密度较低，节奏比较舒缓，整体调式为 C♯小调，代表了情感为 A0-V1 的音乐。

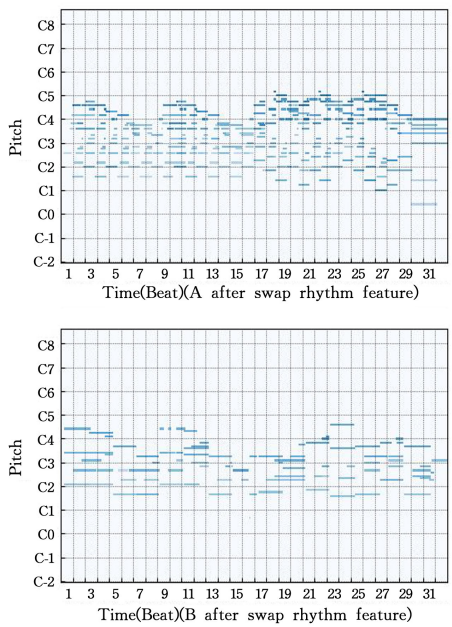
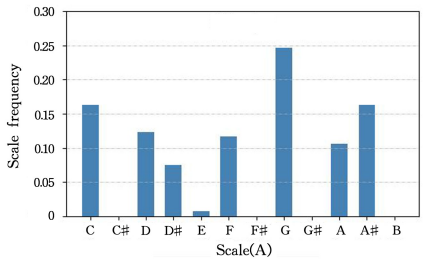
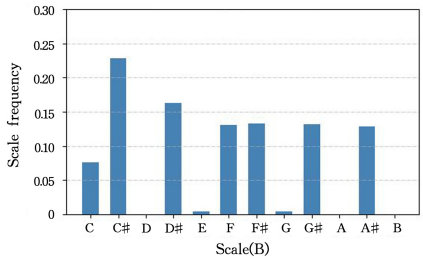


图 13　原始音乐样本 A 和 B 的钢琴卷帘矩阵

Fig. 13　Pianoroll matrix of original music samples A and B



(a)C♯ major



(b)G minor

图 14 原始音乐样本 A 和 B 的音高直方图

Fig. 14 Pitch histograms of original music samples A and B

当交换 A 和 B 的节奏特征表示后，生成了两首新的音乐样本。其钢琴卷帘矩阵和音高直方图如图 15 和图 16 所示。从图中可以发现，当 A 使用了 B 的节奏特征表示后，其生成的音乐也捕捉到了 B 的节奏模式，音符密度变低，持续时间更长，整体起伏变化较小。而 B 的节奏则变得更加跌宕起伏，表现得更加紧凑。与之相反，由于只改变了 A 和 B 的节奏特征，两者的调式并未发生太大改变。因此，A 和 B 在 Arousal 维度上的情感发生了变化。另一方面，当改变了 A 和 B 的调式特征表示后，也可得到类似的实验结果，如图 17 和图 18 所示。两者的节奏模式保持了原有的状态，而调式在互相学习后产生了互换，A 和 B 在 Valence 维度上的情感有了变化。

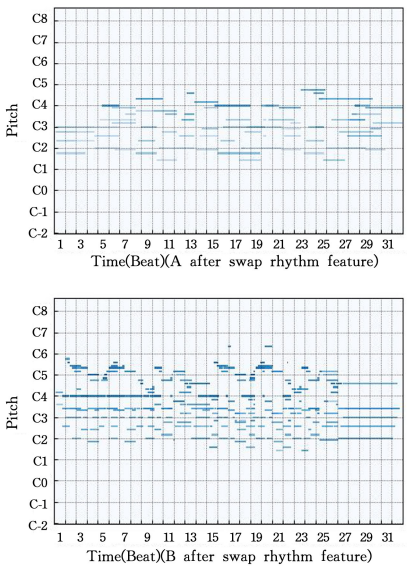
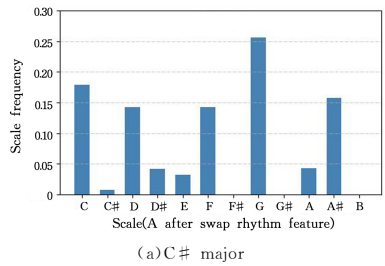
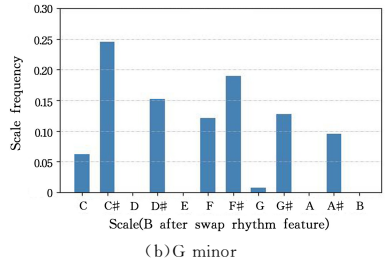


图 15 交换节奏特征表示后音乐样本 A 和 B 的钢琴卷帘矩阵

Fig. 15 Pianoroll matrix of original music samples A and B after swapping rhythm feature representation



(a)C♯ major



(b)G minor

图 16 交换节奏特征表示后音乐样本 A 和 B 的音高直方图

Fig. 16 Pitch histograms of original music samples A and B after swapping rhythm feature representation

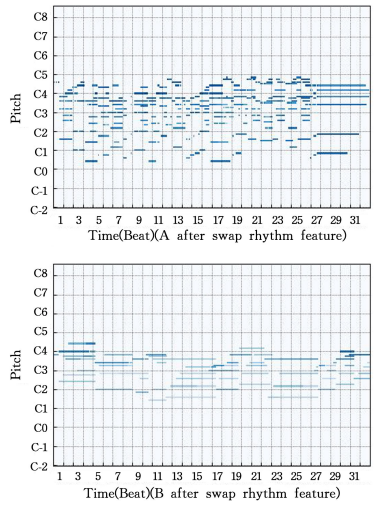
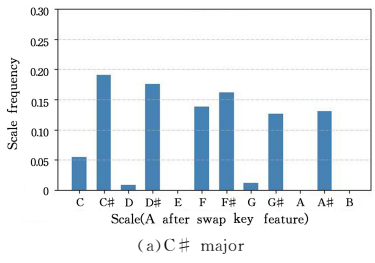
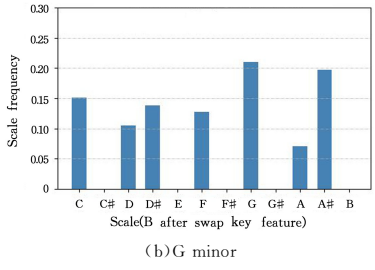


图 17 交换调式特征表示后音乐样本 A 和 B 的钢琴卷帘矩阵

Fig. 17 Pianoroll matrix of original music samples A and B after swapping key feature representation



(a)C♯ major



(b)G minor

图 18 交换调式特征表示后音乐样本 A 和 B 的音高直方图

Fig. 18 Pitch histograms of original music samples A and B after swapping key feature representation

经过以上实验分析，本文提出的情感音乐生成模型 Semg-GMVAE 能够成功从原始音乐序列中分离出与情感相关的节奏特征表示和调式特征表示，并且通过操纵特征表示能够有效实现音乐在 Arousal 和 Valence 维度上的情感变化。

4.4.4　插值情感变换

在上文中，通过交换两首音乐评估了音乐在节奏特征和调式上的解纠缠控制，并且通过这种方法实现了音乐的情感变换。正如 2.2 节所述，交换的变换方法依赖于现有的样本，无法展现生成样本的多样性。为了充分利用 Semg-GMVAE 离散的潜在空间，我们基于特征解纠缠机制和插值法，通过将当前音乐的潜在变量变换到目前情感的潜在空间实现了音乐情感的互相转换。当需要在某一情感维度的高维和低维进行变换时，首先可计算目标情感空间上的高斯分量均值 $\boldsymbol{\mu}_{i,\text{target}}$ 与当前情感空间的高斯分量均值 $\boldsymbol{\mu}_{i,\text{source}}$ 之间的差值，然后将其添加到当前情感的潜在变量 $\boldsymbol{z}_{i,\text{source}}$ 上即可得到目标情感对应的潜在变量 $\boldsymbol{z}_{i,\text{target}}$，最后将该潜在变量输入 Semg-GMVAE 解码器中生成新的样本。整个过程可描述为：

$$\boldsymbol{z}_{i,\text{target}}=\boldsymbol{z}_{i,\text{source}}+\lambda\cdot(\boldsymbol{\mu}_{i,\text{target}}-\boldsymbol{\mu}_{i,\text{source}}) \tag{21}$$

其中，i 代表需要变换的情感维度；参数 $\lambda\in[0,1]$，可控制当前情感与目前情感的紧密程度，在本实验中，为了更明显地看到情感变化，λ 值设为 1。

根据上述方法，我们在测试集进行了音乐情感转换的实验。根据 Arousal 和 Valence 组成的高低维情感空间，将音乐在快乐、紧张、悲伤、平静 4 种情感下进行相互转换，对于最终的结果，同样使用音乐情感分类模型 EMPOIA 进行准确率的评估，得到如表 4 所列的混淆矩阵。

表 4　音乐情感转换准确率评估

Table 4　Evaluation of the accuracy of music emotion transformation

（%）

Source	快乐 (A1-V1)	紧张 (A1-V0)	悲伤 (A0-V0)	平静 (A0-V1)
快乐 (A1-V1)	—	60.84	59.70	**71.56**
紧张 (A1-V0)	63.45	—	**67.81**	57.63
悲伤 (A0-V0)	58.29	**65.48**	—	64.80
平静 (A0-V1)	**70.26**	56.21	62.35	—

从总体上看，通过插值法将潜在变量变换到目标情感类簇上能够实现一定程度的情感转换。这也说明了各个情感高斯分量之间得到了很好的分离，每一种情感通过潜在变量信息得到了准确的表达，但整体的预测准确率并不是很高，这是受到了 EMPOIA 分类模型的影响。在本实验中，EMPOIA 作用在 4 类情感上，而非对单个情感维度的高低维进行预测。实验结果还表明，模型在 Arousal 维度上进行情感转换的准确率较高，在 Valence 维度上进行情感转换的准确率较低，而在两个维度上都进行情感转换的准确率最低。这与 4.4.1 节中得到的结论一致，即音乐在 Arousal 维度上比在 Valence 维度上更容易区分，而同时考虑两个维度上的转换会更加困难。在 4 种情感中，快乐与平静之间的转换准确率达到了最高，这也说明了这两种情感之间存在较高的相似性，仅通过改变节奏模式就能实现较好的转换效果。

4.4.5　主观性实验评估

为进一步验证 Semg-GMVAE 模型的有效性，我们对生成音乐进行了主观性的人工测评。主观性评估实验主要包括两个部分：第一个实验测试生成的音乐是否具备可听性，要求受试者在聆听音乐后针对流畅性、完整性、真实性和情感充分性这 4 方面指标进行评分；第二个实验测试生成的音乐是否具有特定的情感，同样要求受试者针对目标情感符合的程度得出相应评分。所有指标的评分范围设置为 1～5，从低到高分别对应“很差、较差、一般、较好、很好”和“完全不符合、不太符合、一般、较符合、非常符合”5 个等级。在实验过程中，我们准备了 100 首 Semg-GMVAE 模型生成的音乐与 50 首 LPD 数据集中的真实音乐，涉及多种情感的生成和转换，并在大学校园内招募了 10 名音乐爱好者作为受试者，要求每人聆听不同的 10 首生成音乐和 5 首真实音乐并给出评分。生成音乐与真实音乐未事先告知受试者。以真实音乐为基准，对比判断生成音乐的可听性。在进行情感的主观性实验之前，我们对受试者讲解了 Russell 情感模型的含义以理解情感在不同维度上的变化过程，同时让他们聆听了每一种情感的真实音乐以进行情感校准。此外，为了确保实验的准确性和可靠性，每一位受试者在聆听每首音乐后都会休息 3 min，以平复情绪，消除对收听下一首音乐的主观影响。

图 19 展示了第一个主观性评估实验的结果，包含所有生成音乐与真实音乐在流畅性、完整性、真实性和情感充分性 4 种指标上的平均得分。首先，受试者对于真实音乐的反应是正常的，在所有指标上都给出了接近 5 的评分，这有助于受试者对模型生成的音乐作出相对客观的评价。从整体上看，Semg-GMVAE 模型能够有效学习音乐的内在结构，生成的音乐都获得了较高的评分，在听感上接近真实的音乐。在完整性与情感充分性方面，受试者们给出了与真实音乐最接近的两个评分。这一方面得益于 Transformer-XL 对更多音乐小节的上下文结构的把控，另一方面也证实了音乐特征与情感之间的关联赋予了生成音乐更好的情感表达。然而，生成音乐在流畅性方面还存在不足。Semg-GMVAE 模型对音乐节奏变化、重复递进的学习仍有待提高。

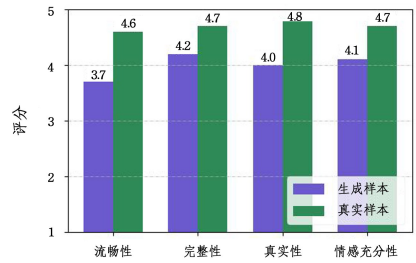


图 19　生成音乐与真实音乐关于可听性的主观评分

Fig. 19　Subjective score of generated musicand real music with respect to audibility

表 5 列出了第二个主观性评估实验的结果，即对特定情感音乐的生成与转换的评分情况。与 4.4.4 小节中客观性评估实验得到的结果类似，在 Arousal 维度上进行特定情感的转换得到了较高的评分，而在 Valence 维度上的评分较低，并且在两个维度上都进行情感转换时得到的评分最低。与客观

实验结果不同的是，受试者对音乐在悲伤与平静之间进行情感转换的评分较高。对于该结果，我们重新对受试者进行了反馈调研，认为悲伤与平静的音乐在听感上较为相似，主观感受较难准确区分，因此都给出了较高的评分。此外，我们测试了从指定情感类别的高斯分量中采样潜在变量生成的效果，对应于表中主对角线上的评分值。可以发现，由于 Semg-GMVAE 模型的各个高斯分量能够正确反映不同情感音乐的特征表示，因此生成的音乐在 4 种情感上都获得了较高的得分，并且快乐和平静的评分最高。这可能是由于受试者在聆听音乐时都具有积极的情感状态，因此倾向于给出积极的情感评分。这说明了音乐对于人类情感的影响并不局限于其本身的情感，人类内心的情感也会与音乐产生共鸣。

表 5 音乐在特定情感上生成和转换的主观评分

Table 5 Subjective score of music generation and transformation on specific emotions

Source	快乐 (A1-V1)	紧张 (A1-V0)	悲伤 (A0-V0)	平静 (A0-V1)
快乐(A1-V1)	**4.6**	3.5	3.4	**4.1**
紧张(A1-V0)	3.6	**4.2**	3.7	3.3
悲伤(A0-V0)	3.4	3.6	**4.4**	**3.9**
平静(A0-V1)	**4.0**	3.2	**3.7**	**4.5**

结束语 本文提出了一种基于改进的高斯混合变分自编码器的情感音乐生成模型(Semg-GMVAE)。该模型通过 GMVAE 在训练过程中以半监督方法进行聚类推断，相较于其他完全监督的生成模型，能够基于少量有标签数据学习更丰富的数据分布，有效缓解了情感音乐数据不足的问题，提升了对于不同情感音乐的类别推断能力与生成能力。针对 GMVAE 存在的模式坍塌问题，即各个高斯分量在训练过程中无法更好地分离，难以区分不同类别数据，影响学习表示的性能和生成的质量，本文进一步分析了 GMVAE 的证据下界，认为其中的方差正则项和互信息抑制项是导致该问题的重要原因。因此本文对这两项因子分别进行惩罚和增强。实验证明该方法能够保证不同情感的音乐在潜在空间得到更好的分离效果，增强了音乐与情感信息的相关程度，提高了半监督模型的鲁棒性和泛化能力。考虑到现有音乐情感生成模型大多直接基于情感标签生成，无法给予情感的可解释性，本文提出了通过音乐的节奏特征和调式特征与情感建立联系，在模型中引入了特征解纠缠机制学习这两种特征的情感表示，加入了生成对抗损失来提升特征解纠缠能力，从而实现了通过操纵特征来控制音乐情感的转换。此外，本文还使用 Transformer-XL 作为 GMVAE 的编码器和解码器，能够有效学习音乐序列中更长的上下文依赖结构，进一步提升了生成音乐的真实性。

然而，本文提出的情感音乐生成模型依然存在一些不足。首先，其对于情感的表达依然不够准确，尤其是在 Valence 维度上的情感推断准确率较低，因此应该寻求能够更准确表达该情感维度的音乐特征。其次，由于情感的主观性，模型应该具备处理标签噪声的能力。在之后的研究中，我们计划使用一致性正则化方法、多标签生成、弱监督学习来改善这一问题。最后，现有的音乐情感生成通常以符号化 MIDI 形式呈现，但音乐情感的表达可以涉及多种感知模态，例如音乐的歌词、音乐的文字评价、演奏人的动作表情、舞台灯光效果等，未来的研究可以尝试将不同的模态进行融合，实现跨模态情感生成，从而创造更具丰富度和表现力的音乐情感体验。

参 考 文 献

[1] TIE Y, CHEN H J, JIN C, et al. Research on emotion recognition method based on audio and video feature fusion[J]. Journal of Chongqing University of Technology(Natural Science), 2022, 36(1): 120-127.

[2] MA L, ZHONG W, MA X, et al. Learning to generate emotional music correlated with music structure features[J]. Cognitive Computation and Systems, 2022, 4(2): 100-107.

[3] SULUN S, DAVIES M E P, VIANA P. Symbolic music generation conditioned on continuous-valued emotions[J]. IEEE Access, 2022, 10: 44617-44626.

[4] HUNG H T, CHING J, DOH S, et al. EMOPIA: A Multi-Modal Pop Piano Dataset For Emotion Recognition and Emotion-based Music Generation[C]// Proceedings of 22th International Society for Music Information Retrieval Conference(ISMIR). 2021: 318-325.

[5] KINGMA D P, WELLING M. Auto-encoding variational bayes[J]. arXiv: 1312.6114, 2013.

[6] GREKOW J, DIMITROVA-GREKOW T. Monophonic music generation with a given emotion using conditional variational autoencoder[J]. IEEE Access, 2021, 9: 129088-129101.

[7] TAN H H, HERREMANS D. Music FaderNets: Controllable Music Generation Based on High-Level Features via Low-Level Feature Modelling[C]// Proceedings of 21th International Society for Music Information Retrieval Conference (ISMIR). 2020: 109-116.

[8] DILOKTHANAKUL N, MEDIANO P A M, GARNELO M, et al. Deep unsupervised clustering with Gaussian mixture variational autoencoders[C]// International Conference on Learning Representations(ICLR). 2017.

[9] RUSSELL J A. A circumplex model of affect[J]. Journal of Personality and Social Psychology, 1980, 39(6): 1161.

[10] LI Z, ZHAO Y, XU H, et al. Unsupervised clustering through Gaussian mixture variational autoencoder with non-reparameterized variational inference and std annealing[C]// 2020 International Joint Conference on Neural Networks(IJCNN). IEEE, 2020: 1-8.

[11] ROBERTS A, ENGEL J, RAFFEL C, et al. Ahierarchical latent vector model for learning long-term structure in music[C]// International Conference on Machine Learning(ICML). PMLR, 2018: 4364-4373.

[12] BRUNNER G, KONRAD A, WANG Y, et al. MIDI-VAE: Modeling dynamics and instrumentation of music with applications to style transfer[C]// Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR). 2018: 747-754.

[13] DAI Z, YANG Z, YANG Y, et al. Transformer-XL: Attentive Language Models beyond a Fixed-Length Context[C]// Proceedings of the 57thAnnual Meeting of the Association for Com-

putational Linguistics(ACL). 2019:2978-2988.

[14] HEVNER K. Experimental studies of the elements of expression in music[J]. The American Journal of Psychology,1936,48(2):246-268.

[15] CHEŁKOWSKA-ZACHAREWICZ M,JANOWSKI M. Polish adaptation of the Geneva Emotional Music Scale:Factor structure and reliability[J]. Psychology of Music,2021,49(5):1117-1131.

[16] THAYER R E. The biopsychology of mood and arousal[M]. Oxford University Press,1990.

[17] MEHRABIAN A. Silent messages:implicit communication of emotions and attitudes[M]. Wadsworth Pub,1981.

[18] KREUTZ G,OTT U,TEICHMANN D,et al. Using music to induce emotions:Influences of musical preference and absorption [J]. Psychology of Music,2008,36(1):101-126.

[19] VIEILLARD S,PERETZ I,GOSSELIN N,et al. Happy,sad,scary and peaceful musical excerpts for research on emotions [J]. Cognition & Emotion,2008,22(4):720-752.

[20] YANG X,SONG Z,KING I,et al. A survey on deep semi-supervised learning[J]. arXiv:2103.00550,2021.

[21] KINGMA D P,REZENDE D J,MOHAMED S,et al. Semi-supervised learning with deep generative models[C]// Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 2(NIPS). 2014:3581-3589.

[22] HABIB R,MARIOORYAD S,SHANNON M,et al. Semi-supervised generative modeling for controllable speech synthesis [C] // International Conference on Learning Representations (ICLR). 2019.

[23] CHEUNG V K M,KAO H K,SU L. Semi-supervised violin fingering generation using variational autoencoders[C] // Proceedings of 22th International Society for Music Information Retrieval Conference(ISMIR). 2021:113-120.

[24] SCHUSTER M,PALIWAL K K. Bidirectional recurrent neural networks[J]. IEEE Transactions on Signal Processing,1997,45(11):2673-2681.

[25] LI Y,PAN Q,WANG S,et al. Disentangled variational autoencoder for semi-supervised learning[J]. Information Sciences,2019,482:73-85.

[26] JOY T,SCHMON S M,TORR P H S,et al. Capturing label characteristics in VAEs[C] // International Conference on Learning Representations(ICLR). 2021.

[27] DEMPSTER A P,LAIRD N M,RUBIN D B. Maximum likelihood from incomplete data via the EM algorithm[J]. Journal of the Royal Statistical Society:Series B(Methodological),1977,39(1):1-22.

[28] LUO Y J,AGRES K,HERREMANS D. Learning disentangled representations of timbre and pitch for musical instrument sounds using Gaussian mixture variational autoencoders[C]// Proceedings of 20th International Society for Music Information Retrieval Conference(ISMIR). 2019:746-753.

[29] BENGIO Y,COURVILLE A,VINCENT P. Representation learning:A review and new perspectives[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2013,35(8):1798-1828.

[30] WANG X,CHEN H,TANG S,et al. Disentangled Representation Learning[J]. arXiv:2211.11695,2022.

[31] HIGGINS I,MATTHEY L,PALA,et al. beta-VAE:Learning basic visual concepts with a constrained variational framework [C] // International Conference on Learning Representations (ICLR). 2017.

[32] CHEN R T Q,LI X,GROSSE R,et al. Isolating sources of disentanglement inVAEs[C]// Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS). 2018:2615-2625.

[33] KUMAR A,SATTIGERI P,BALAKRISHNAN A. Variational inference of disentangled latent concepts from unlabeled observations[C]// International Conference on Learning Representations(ICLR). 2018.

[34] WANG Z,WANG D,ZHANG Y,et al. Learning interpretable representation for controllable polyphonic music generation [C]// Proceedings of 21th International Society for Music Information Retrieval Conference(ISMIR). 2020:662-669.

[35] YANG R,WANG D,WANG Z,et al. Deep music analogy via latent representation disentanglement[C] // Proceedings of 20th International Society for Music Information Retrieval Conference(ISMIR). 2019:596-603.

[36] WU Y,CARSAULT T,NAKAMURA E,et al. Semi-supervised neural chord estimation based on a variational autoencoder with latent chord labels and features[J]. IEEE/ACM Transactions on Audio,Speech,and Language Processing,2020,28:2956-2966.

[37] AKAMA T. Controlling Symbolic Music Generation based on Concept Learning from Domain Knowledge[C]// Proceedings of 20th International Society for Music Information Retrieval Conference(ISMIR). 2019:816-823.

[38] CHOI K,CHO K. Deep unsupervised drum transcription[C]// Proceedings of 20th International Society for Music Information Retrieval Conference(ISMIR). 2019:183-191.

[39] ZHANG Y. Representation learning for controllable music generation:A survey[C]// Proceedings of 20th International Society for Music Information Retrieval Conference (ISMIR). 2020:1-8.

[40] MI L,HE T,PARK C F,et al. Revisiting LatentSpace Interpolation via a Quantitative Evaluation Framework[J]. arXiv:2110.06421,2021.

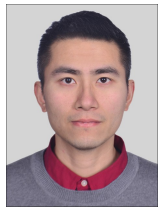
[41] JIANG Z,ZHENG Y,TAN H,et al. Variational deep embedding:an unsupervised and generative approach to clustering [C]// Proceedings of the 26th International Joint Conference on Artificial Intelligence(IJCAI). 2017:1965-1972.

[42] ZHAO T,LEE K,ESKENAZI M. Unsupervised Discrete Sentence Representation Learning for Interpretable Neural Dialog Generation[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics(ACL). 2018:1098-1107.

[43] REZAABAD A L,VISHWANATH S. Learning representations by maximizing mutual information in variational autoencoders [C] // 2020 IEEE International Symposium on Information Theory(ISIT). IEEE,2020:2729-2734.

[44] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is

all you need[C]// Advances in Neural Information Processing Systems(NeurIPS). 2017:5998-6008.
[45] JIANG J, XIA G G, CARLTON D B, et al. Transformer VAE: A Hierarchical Model for Structure-Aware and Interpretable Music Representation Learning[C]// 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020). IEEE, 2020:516-520.
[46] WU S L, YANG Y H. MuseMorphose: Full-song and fine-grained piano music style transfer with one transformer VAE [J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31:1953-1967.
[47] DONG H W, HSIAO W Y, YANG L C, et al. MuseGAN: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment[C]// Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). 2018: 34-41.
[48] BERTIN-MAHIEUX T, ELLIS D P W, WHITMAN B, et al. The million song dataset[C]// Proceedings of 12th International Society for Music Information Retrieval Conference (ISMIR). 2011:591-596.
[49] HUANG Y S, YANG Y H. Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions [C]// Proceedings of the 28th ACM International Conference on Multimedia (ACM Multimedia). 2020:1180-1188.
[50] GLOROT X, BENGIO Y. Understanding the difficulty of training deep feedforward neural networks[C]// Proceedings of the 13th International Conference on Artificial Intelligence and Statistics(AISTATS). JMLR Workshop and Conference Proceedings, 2010:249-256.
[51] ZHENG K, MENG R, ZHENG C, et al. EmotionBox: A music-element-driven emotional music generation system based on music psychology[J]. Frontiers in Psychology, 2022, 13:5189.
[52] VAN DER MAATEN L, HINTON G. Visualizing Data using t-SNE[J]. Journal of Machine Learning Research, 2008, 9:2579-2605.
[53] KAWAI L, ESLING P, HARADA T. Attributes-Aware Deep Music Transformation[C]// Proceedings of 21th International Society for Music Information Retrieval Conference (ISMIR). 2020:670-677.
[54] DONG H W, HSIAO W Y, YANG Y H. Pypianoroll: Open source Python package for handling multitrack pianorolls[C]// Proceedings of 19th International Society for Music Information Retrieval Conference(ISMIR). Late-breaking paper, 2018.



XU Bei, born in 1986, Ph. D, associate professor, is a member of CCF (No. P1014M). His main research interests include affective computing and natural language processing.

LIU Tong, born in 1997, postgraduate. His main research interests include affective computing and music generation.

(责任编辑:何杨)