文章编号：2095-2163(2022)09-0110-05 中图分类号：TP391 文献标志码：A

基于多模态深度学习的音乐情感分类算法

周 萍

（南昌职业大学 信息技术学院，南昌 330500）

摘 要：针对现有算法仅考虑音乐或视频单模态特征，分类效率低下的问题，本文提出了基于多模态深度学习的音乐情感分类算法。首先设计了二维音频卷积神经网络，该网络将音频梅尔谱图作为输入，以学习音乐的音频特征；其次，设计了视频神经网络，以学习音乐视频的时空特征；采用多模态融合技术，将音频和视频特征融合，设计了多模态深度学习分类算法对音乐的情感进行分类。针对缺乏已标记音乐视频数据集的问题，本文构建了一个具有多样性的音乐视频数据集。基于该数据集进行实验，以验证所提出算法的有效性，并对比分析了不同优化器对单模态分类模型性能的影响。实验结果表明，与单模态情感分类器相比，多模态分类器能实现最佳的分类性能。

关键词：深度学习；音乐情感分类；多模态融合

Music emotion classification algorithm based on multimodal deep learning

ZHOU Ping

（School of Information Technology, Nanchang Vocational University, Nanchang 330500, China）

【Abstract】 Aiming at the problem that the existing algorithms only consider the single-modal features of music or video and the classification efficiency is low, this paper proposes a music emotion classification algorithm based on multi-modal deep learning. First, this paper proposes a two-dimension audio convolutional neural network that takes audio mel spectrograms as input to learn audio features of music; after that, this paper designs a video neural network to learn the spatiotemporal features of music videos. Then, this paper uses multimodal integration technology to fuse audio and video features, and designs a multimodal deep learning classification algorithm to classify the emotion of music.Finally, in response to the lack of labeled music video datasets, this research constructs a diverse music video dataset. This paper conducts experiments based on this data set to verify the effectiveness of the proposed algorithm, and compares and analyzes the effects of different optimizers on the performance of the single-modal classification model. The experimental results show that the multimodal classifier achieves the best classification performance compared to each unimodal sentiment classifier.

【Key words】 deep learning; music emotion classification; multimodal fusion

0 引 言

众所皆知，音乐艺术家通常使用动态节奏、发音来传达音乐中的情感。随着互联网和流视频技术的发展，“音乐+视频”逐渐成为流行的可视化表现形式。对于音乐视频，用户在关注其名称、收录专辑和艺术家的同时，还会关注诸如流派、情感和视频质量等属性。因此，结合视频内容对音乐的情感等属性进行分类是一个重要的研究课题，亦能切合线上音乐网站、音乐视频网站和内容共享网络等场景的应用需求。现有的研究主要集中于分别面向音频和视频单模态信息来进行分类等任务，但是针对音乐视频进行情感分析仍然是一个亟待解决的热点问题。情感可以通过情感词汇以口头方式表达，也可以通过非语言线索（如语调、面部表情和手势）表达。音乐视频中的情感不仅包括了视频、文字、面部表情等情感属性，还涵盖了通过音乐旋律、器乐节奏和作曲家突出场景表达的附加情感。本文将采用深度学习算法对音乐视频的情感进行分类，提出以音频梅尔谱图为输入的音频神经网络以学习音频特征，剖析了视频神经网络学习视频数据的时空特征，用于捕获整个视频信息。本文使用多模态融合，将音频特征和视频特征结合进行情感分类。由于缺乏已标记音乐视频数据集，本文构建了具有多样性的音乐视频数据集，并基于该数据集进行实验评估，用来验证提出的算法的有效性。

1 音频和视频神经网络设计

1.1 音频神经网络设计

卷积神经网络（Convolutional Neural Network，CNN）的有效性在于能够从端到端管道中的原始数据中学习特征以应对特定任务。由于处理诸如音频

作者简介：周 萍（1980-），女，硕士，讲师，主要研究方向：计算机及应用研究。

收稿日期：2022-03-06

这一类信号需要二维 CNN,因此本文设计了二维音频网络来提取音频特征。许多现有的音频网络使用音乐的梅尔谱图的幅度表示作为输入,忽略了梅尔谱图的相位信息。音频网络使用梅尔谱图的原始波形作为输入可以同时保留信号幅度和相位信息,因此本文提出的二维音频网络需要使用原始波形作为输入进行音乐情感分析。本文提出的音频网络结构如图 1 所示。由图 1 中的二维音频网络可知,卷积层(32)是指在池化大小为 2 的通道上执行 32 个步长为 2 卷积操作,卷积核的大小为 3×1,再采用最大池化操作来降低信号的维数,同时保留卷积信号中的必要统计信息。本文提出的二维音频网络使用指数线性单元(*ELU*)作为激活函数[1],使用 *Dropout*(0.1)缓解过拟合的副作用[2],其中 *Dropout*(0.1)是指概率为 0.1 的 *dropout* 操作。本文提出的二维音频网络将来自 4 个卷积层和 4 个最大池化层的输出拼接为一个 32 × 8 的二维输出矩阵。

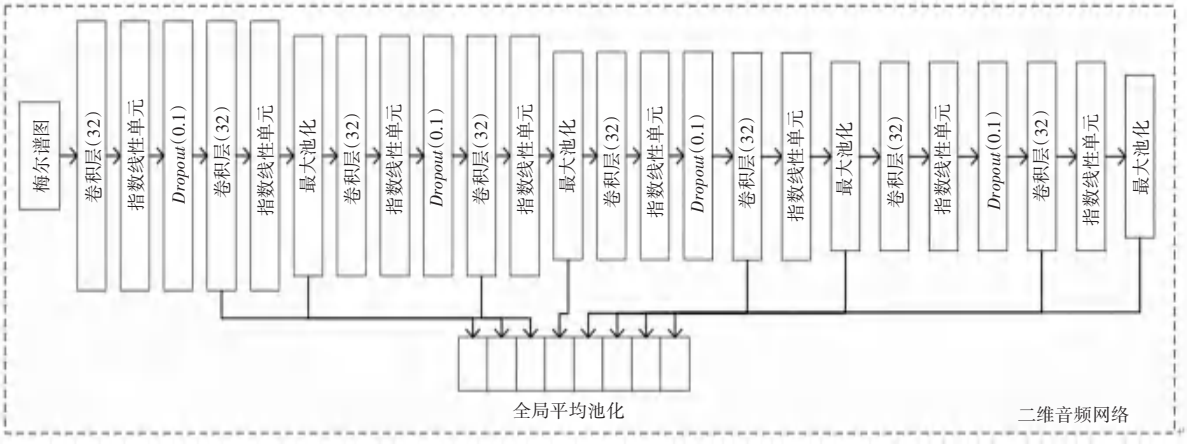

图 1　二维音频网络

Fig. 1　Two-dimension audio network

1.2　视频神经网络设计

本文采用三维卷积网络(3D Convolutional Networks, C3D)和膨胀卷积网络(Inflated 3D ConvNet,I3D)对视频特征进行提取。二维音频网络和 C3D 视频网络融合流程如图 2 所示。由图 2 可见,在进行 C3D 预处理后,本文将原始 C3D 网络的末端 2 个全连接层替换为维度分别为 1 024 和 512 的全连接层,以降低维度,并应用概率为 0.2 的 *Dropout* 层来缓解音乐视频数据微调中的过度拟合问题。使用随机梯度下降(SGD)作为优化器,学习率设置为 0.000 01。此外,研究还给出了二维音频网络和 I3D 网络视频融合流程如图 3 所示。图 3 中,对于 I3D 网络,对最后的 inception 块的输出进行三维全局平均池化,使用音乐视频数据对整个网络进行微调,并使用学习率为0.000 1的 Adam 优化器。

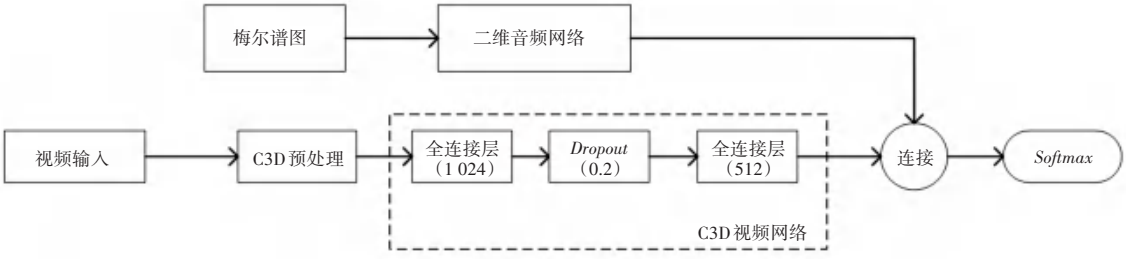

图 2　二维音频网络和 C3D 视频网络融合

Fig. 2　The fusion between two-dimension audio network and C3D video network

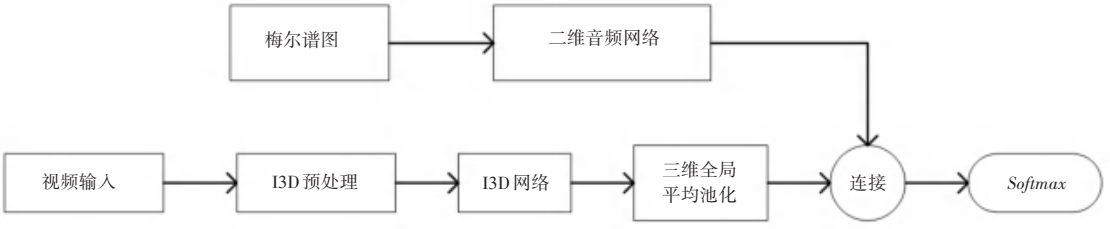

图 3 二维音频网络和 I3D 网络视频融合

Fig. 3 The fusion between two-dimension audio network and I3D video network

2 分类算法设计

2.1 输入预处理

在使用多模态深度学习算法进行训练前,需要对输入的视频数据进行预处理。图 2、图 3 中的 C3D 预处理和 I3D 预处理过程即如图 4 所示。使用 C3D 和 I3D 视频网络进行视觉情感分类之前,所有的视频帧被调整到合适的大小、数量和通道。C3D 和 I3D 网络使用具有红色、绿色和蓝色通道的 32 帧视频进行训练。对于视频帧,以统一的时间间隔进行提取以捕获整个视频信息内容。图 4 中,卷积层 1 包含 64 个卷积操作,卷积层 2 包含 256 个卷积操作,卷积层 3 包含 512 个卷积操作,卷积层 4 包含一个大小为 7×7×7、步长为 2 的卷积核,卷积层 5 包含一个大小为 1×1×1 的卷积核,卷积层 6 包含一个大小为 3×3×3 的卷积核。对于音频数据,本文对音频数据进行零填充得到全长音频波形,随后对 30 s 的音乐信号以 16 kHz 的频率进行采样。音频输入是 CNN 音乐网络的 480 000 维输入向量。经过预训练的 2D 音乐 CNN 也需要使用零填充生成全长音频,以生成固定大小的梅尔谱图。梅尔谱图是通过获取短时傅里叶变换(short-time Fourier Transform, STFT)的绝对值发现的频率内容随时间的二维表示。

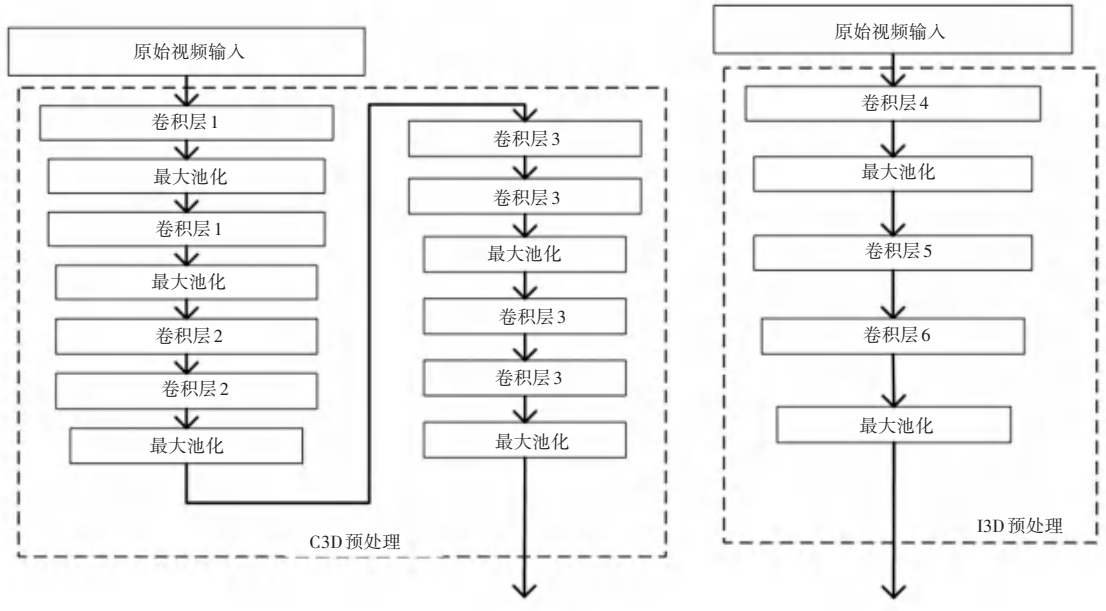

图 4 输入预处理

Fig. 4 Input preprocessing

2.2 多模态分类

将音频和视频信息融合为多模态,分别将 C3D 和 I3D 视频网络的决策级特征与一维音乐 CNN(OneDMCNN)和二维音乐 CNN(TwoDMCNN)的决策级特征融合,产生了 4 种多模态架构,即 C3D + OneDMCNN、I3D + OneDMCNN、C3D + TwoDMCNN 和 I3D + TwoDMCNN。音频和视频的每个单模态情感分类器首先分别使用数据集进行微调,去除每个单模态的 *SoftMax* 分类器后,将输出的结果用于多模态特征融合。

为了克服数据匮乏的问题,结合迁移学习来进行音乐视频分类。首先加载预训练的权重,并微调

源神经网络，使其适应音乐视频数据集。然后，提取每个单模态情感分类的学习特征，用于多模态决策。使用 sport－1 M 数据集训练 C3D[3] 以及使用 RGB ImageNet 和 kinetic 数据集[4]训练 I3D。采用歌曲数据集训练的预训练二维音乐 CNN 作为音乐情感分类器，并微调该音频网络，以对网络分类音乐情感进行泛化。

多模态融合是一个整合来自多个来源信息的过程，目前有 3 种信息融合方法，包括早期融合、晚期融合和混合融合。应用了后期融合，将最高级别的预训练特征组合起来，由 *SoftMax* 层做出最终分类决策。将每个单模态网络所学习的特征连接起来，用于单独的音乐和视频情感决策，也由 *SoftMax* 层做出最终决策。

3　实验评估

3.1　数据集构建

现有的情感分类算法应用机器学习技术来训练分类器，将情感划分为离散的类别。这些算法通过训练预测输入数据的情感类别，从而将输入表示为情感空间中的一个点。

音乐视频情感分析的主要挑战在于情感边界的模糊性和标记训练数据的稀缺性。现有的音乐视频数据集仍无法用于情感分类算法的有效训练，而数据集的稀缺问题对研究人员来说是一个巨大的挑战，算法需要足够的数据样本才能做出更准确的决策。对此，首先通过整合现有的数据集[5-6]和其他从互联网上收集的数据样本，构建了一个用于音乐视频情感分析的小型数据集。将多种人类情感划分为 6 个类别作为基本情感类别，即兴奋、恐惧、中性、放松、悲伤和紧张。上述属于 6 个情感类别的样本如图 5 所示，每个音乐视频样本的长度约为 30 s。

图 5　数据集情感实例

Fig. 5　Emotion examples of the dataset

在本音乐视频数据集中，大多数音乐视频是从互联网上收集的，这使得数据集在区域、语言、文化和乐器方面存在巨大差异。每个数据样本都有其不同的特征，包括频率、音高、过零率、运动强度、节奏规律和分辨率等。本文所考虑的 6 种情感类别的边界是模糊的，部分情感之间存在重叠。选择了一致且易于确定情感的音乐视频来构建数据集，并用对应于 6 种基本类别的情感对数据集进行标注。其中，兴奋的情绪通常是指人们高兴或受某种刺激而精神激奋，视觉内容一般包括舞蹈或派对等高强度肢体动作、群体活动和丰富多变的环境。这类音乐一般采用快节奏、大调、和声、流畅或多变的节奏。恐惧情绪源于对危险或恐怖的感知，视觉信息包括一些不自然的事件或随时间突然变化的人物，音乐创作时通常会用到快节奏、高响度和不规则的节奏。放松是一种低张力状态或恢复平衡状态。这一类的视觉信息一般包括自然场景和乐器，此类别的音乐通常具有缓慢的节奏与和声。悲伤是人类心理的一种不愉快的感觉，悲伤的音乐通常具有缓慢的节奏和轻微的音调。紧张类别包括引发负面情绪的暴力场景，带有紧张情绪的音乐一般包括高响度、快节奏和碰撞的和声。

3.2　实验设置和结果分析

在实验部分，分别使用音频和视频神经网络来测试提出的音乐视频数据集。使用迁移学习并对预训练的 CNN 进行微调，以将其应用于提出的音乐视频数据集。实验中使用的性能评估指标主要有：准确率、F_1 分数和受试者操作特征曲线下的面积（*AUC*）。其中，准确率是指正确分类的数据样本占总样本的百分比，F_1 分数是精度和召回率之间的调和平均值。受试者操作特征曲线是显示分类算法在所有分类阈值下的性能以及真假阳性率的图表，而 *AUC* 是曲线下方的整个二维区域，表示了对所有可能分类阈值的性能的聚合度量。针对不同的神经网络，实验选择了 2 种优化器。优化器的作用是用于调整神经网络的参数，使神经网络更快、更好地收敛。表 1 展示了各种单模态分类网络的评估结果以及优化器对各种学习因素的影响。由结果可知，一维音乐 CNN 和 I3D 使用 Adam 优化器时能获得最好的性能，二维音乐 CNN 和 C3D 使用 SGD 优化器时能获得最好的性能。学习率设置为 0.001。

随后将音频和视频的单模态结果融合为最终的多模态结构。C3D 使用 Adam 和 SGD 优化器的性能较为接近，本文仅选择 SGD 优化器进行多模态集成。二维音乐 CNN 在歌曲数据集上进行预训练，因

此其性能优于一维音乐 CNN。虽然一维音乐 CNN 包含了音频流的相位和幅度,但由于端到端训练的数据样本非常有限,一维音乐 CNN 的性能也无法超过二维音乐 CNN。

表 1 不同优化器下单模态分类网络的评估结果

Tab. 1 Evaluation results of single - modality classification networks under different optimizers

优化器	指标	OneDMCNN	TwoDMCNN	C3D	I3D
Adam	准确率	0.45	0.65	0.64	0.66
	F_1 分数	0.41	0.65	0.64	0.66
	AUC	0.76	0.92	0.90	0.90
SGD	准确率	0.42	0.73	0.64	0.60
	F_1 分数	0.37	0.72	0.64	0.59
	AUC	0.73	0.93	0.89	0.88

将表现最好的单模态分类器中学习到的特征整合到各种优化器上,用于音乐视频情感预测。音乐网络在决策级别与视频网络集成,并使用 *SoftMax* 分类器对级联特征进行分类。分类是通过六重交叉验证完成的。表 2 展示了各种多模态组合的结果。决策级特征融合的结果表明,当所有的音视频特征都用 *SoftMax* 决策算子结合时,能获得最好的性能。为了更好地了解提出算法的性能,实验统计了所有多模态分类器及集成多模态的 *AUC*。由表 2 结果可知,集成多模态在决策级别融合了多模态的所有学习特征,因此拥有最好的 *AUC* 性能。集成多模态的混淆矩阵见表 3。由表 3 结果可知,与其他情绪相比,放松和悲伤情绪更容易被混淆。与其他无声情绪相比,兴奋和紧张情绪之间能够得到更好的区分。

表 2 单模态组合和集成多模态的评估结果

Tab. 2 Evaluation results of single - modality combination and integrated multi-modality

指标	C3D + OneDMCNN	I3D + OneDMCNN	C3D + TwoDMCNN	I3D + TwoDMCNN	集成多模态
准确率	0.73	0.70	0.85	0.84	0.86
F_1 分数	0.70	0.69	0.84	0.84	0.88
AUC	0.92	0.93	0.98	0.98	0.99

表 3 集成多模态的混淆矩阵

Tab. 3 Confusion matrix for integrated multi-modality

	兴奋	恐惧	中性	放松	悲伤	紧张
兴奋	72	3	7	0	5	0
恐惧	4	68	6	3	0	0
中性	2	3	69	3	4	0
放松	2	1	3	85	7	1
悲伤	3	1	8	7	63	3
紧张	2	1	1	1	3	62

4 结束语

借助迁移学习和后期决策级融合,提出了基于多模态深度学习的音乐视频情感分类算法。构建了一个小型音乐视频数据集,将音乐和视频部分分开,以便用于其他音频和视频 CNN 的预训练。实验评估的结果表明,多模态融合能有效提高分类性能。该结果表明,在已标记数据样本不足的情况下,提出的算法可以学习到音乐视频的多模态特征,实现准确、高效的情绪分类。

参考文献

[1] KATHAROPOULOS A, VYAS A, PAPPAS N, et al. Transformers are rnns: Fast autoregressive transformers with linear attention[C]// International Conference on Machine Learning. PMLR, 2020: 5156-5165.

[2] TAN Mingxing, LE Q. Efficientnet: Rethinking model scaling for convolutional neural networks[C]//International Conference on Machine Learning. Long Beach, CA, USA: PMLR, 2019: 6105-6114.

[3] ZHANG Hao, HAO Yanbin, NGO C W. Token shift transformer for video classification[C]//Proceedings of the 29th ACM International Conference on Multimedia. ACM, 2021: 917-925.

[4] MOON G, KWON H, LEE K M, et al. Integralaction: Pose-driven feature integration for robust human action recognition in videos[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2021: 3339-3348.

[5] HONG S, IM W, YANG H S. Cbvmr: content-based video-music retrieval using soft intra-modal structure constraint[C]// Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval. Yokohama, Japan: ACM, 2018: 353-361.

[6] KOELSTRA S, MUHL C, SOLEYMANI M, et al. Deap: A database for emotion analysis; using physiological signals[J]. IEEE Transactions on Affective Computing, 2011, 3(1): 18-31.