基于语义嵌入学习的特类视频识别

吴晓雨[1,2],蒲禹江[1],王生进[3],刘子豪[1]
(1. 中国传媒大学信息与通信工程学院,北京 100024;2. 媒体融合与传播国家重点实验室(中国传媒大学),北京 100024;
3. 清华大学电子工程系,北京 100084)

摘 要: 暴力视频传播已经成为网络环境治理面临的隐患之一,暴力视频这类特类视频的智能识别技术对维护互联网内容安全具有重要意义. 由于采集来源的多样性,暴力视频分布通常呈现较大的类内方差和较小的类间方差,常见的暴力视频识别模型难以适应复杂多变的暴力场景. 同时,暴力一词本身具有高度抽象的语义,如何从有限数据中学习通用的暴力语义表示成为一大难点. 针对这些问题,本文基于语义嵌入学习的思想,构建了一种新颖的多模态暴力视频识别模型,主要由三部分构成. (1)多模态特征提取. 考虑到视频具有多模态属性,采用了三种不同的深度神经网络分别提取表观、运动、音频三种模态的特征表示. (2)多模态特征融合. 为获得鲁棒的通用视频表示,设计了一种轻量级的多模态特征融合模块(Multimodal Efficient Fusion Module,MEFM),该模块包括共享空间映射与多模态特征交互两部分,在对多模态特征进行充分交互的同时,又能够有效抑制不同模态信息之间的干扰. (3)语义嵌入学习. 为适应不同数据分布的暴力数据集,提出了一种基于语义嵌入的多任务学习方法,通过引入中心损失构建暴力语义中心,并采用余弦嵌入损失将暴力样本向中心聚合、非暴力样本进行离散,形成具有语义判别性的特征表示,从而增强了模型的泛化能力,减少了数据噪声的干扰. 在 VSD2015,Violent Flows 和 RWF-2000 三个公开数据集上的实验表明,本文提出的暴力视频识别模型较已有方法分别提升了 4.79%,0.81% 和 1.5%,取得了具有竞争力的结果.

关键词: 暴力视频识别;多模态特征融合;语义嵌入;多任务学习
基金项目: 国家自然科学基金(No.61801441)
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112(2023)11-3225-13
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20220601

Special Video Recognition Based on Semantic Embedding Learning

WU Xiao-yu[1,2], PU Yu-jiang[1], WANG Sheng-jin[3], LIU Zi-hao[1]
(1. *School of Information and Communication, Communication University of China, Beijing* 100024, *China*;
2. *State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing* 100024, *China*;
3. *Department of Electronic Engineering, Tsinghua University, Beijing* 100084, *China*)

Abstract: As special type of videos, violent video dissemination has become one of the hidden dangers facing the Internet environment, and intelligent recognition technology for violent videos is of great significance for maintaining Internet content security. Due to the diversity of collection sources, the distribution of violent videos usually shows large intra-class variance and small inter-class variance, and it is difficult for common violence recognition frameworks to adapt to complex and variable violent scenarios. Meanwhile, the word violence itself has highly abstract semantics, and it becomes a major difficulty to learn a generic semantic representation of violence from limited data. In response to these problems, we present a novel multimodal violent video recognition model based on semantic embedding learning. The model mainly consists of the following three parts. (1) Multimodal feature extraction. Considering that videos have multimodal properties, we use three different deep neural networks to extract feature representations of three modalities, i.e., appearance, motion, and audio. (2) Multimodal feature fusion. To obtain a robust universal video representation, a lightweight multimodal feature fusion module, referred to as MEFM (Multimodal Efficient Fusion Module), is designed in this paper. The module includes two parts: common space mapping and multimodal feature interaction, which can effectively suppress the interference between different modal information while fully interacting with multimodal features. (3) Semantic embedding learning. To accommodate violence datasets from different sources, we propose a multi-task learning method based on semantic embed-

收稿日期:2022-05-23;修回日期:2022-08-16;责任编辑:朱梅玉

ding, which computes the semantic center of violence by introducing a center loss and uses cosine embedding loss to aggregate violent samples toward the center while discrete with non-violent samples to form a semantic discriminative feature representation, thus enhancing the generalization ability of the model and reducing the noise interference. Experiments on three publicly available datasets, VSD2015, Violent Flows, and RWF-2000, demonstrate that the violence video recognition framework proposed in this paper achieves competitive results by improving 4.79%, 0.81%, and 1.5% respectively, over the state of the arts.

Key words: violent video recognition; multimodal feature fusion; semantic embedding; multi-task learning
Foundation Item(s): National Natural Science Foundation of China (No.61801441)

1 引言

随着智能移动终端与5G网络的广泛普及与应用,网络视频数量呈现爆发式增长,对视频内容审查的需求也与日俱增.含有血腥、打斗、虐待等内容的暴力视频会对观看者带来视觉和心理上的负面冲击,对构建绿色健康的网络环境产生不良影响.依靠人工的暴力内容审查需要消耗大量的人力物力,不能满足当代海量视频的审核需求[1].因此,暴力视频内容的智能化审查对维护社会和谐稳定、提高网络环境治理具有重要的现实意义.

暴力视频场景复杂且具有敏感性,已有公开数据集规模都不大,这在一定程度增加了研究的难度.在算法研究上,暴力视频识别作为动作识别的一个分支,既包含视频分类的一般特点,又有着自身特性.基于手工设计特征的传统暴力视频识别方法由于受到特征语义描述的限制,识别结果较差.许多研究人员尝试引入深度学习的方法来解决暴力视频识别任务[2~6].文献[3]利用卷积长短时记忆网络提取视频时序特征来检测复杂的视觉暴力行为.文献[4]利用卷积神经网络提取暴力视频视觉通道的时空特征和音频特征,并采用简单的特征拼接方式完成暴力视频的判别,文献[5]基于伪三维卷积网络和长短时记忆网络构建视频时空表示,并引入语义一致性度量作为辅助任务进行暴力识别.文献[6]以捕捉有别于背景的移动对象的背景抑制帧和帧差作为输入,使用可分离卷积长短时记忆网络进行时空编码,以此为基础搭建了一个双流网络来识别暴力行为.

上述研究促进了暴力视频识别领域的研究发展,但是仍存在两个关键问题.(1)多模态特征融合过程简单粗暴.目前,基于多模态特征融合的暴力视频识别模型通常使用特征拼接或相加的方法,抑或采用简单的线性层映射加权,未能较好地平衡多模态的互补性和一致性,融合效果有限.(2)未充分考虑暴力数据集的分布特性.与常见的分类任务不同,“暴力”本身是高度抽象的语义概念,其中包括血腥、打斗、枪击、爆炸等具体的暴力元素.因此,暴力视频数据通常呈现较大的类内方差和较小的类间方差,这对暴力识别模型的通用性提出了一定挑战.

针对现有方法的不足,本文提出了一个语义嵌入学习的多模态暴力视频识别方法.首先,在暴力视频的表观、运动、音频多模态特征提取基础上,设计了一种多模态特征高效融合模块(Multimodal Efficient Fusion Module,MEFM),通过共享空间映射和多模态特征交互两部分较好地兼容了多模态特征的互补性和一致性,以解决暴力视频多模态特征无法有效融合影响识别性能的问题;其次,提出了一种新颖的语义嵌入学习策略,通过引入中心损失构建暴力样本的语义中心以聚合暴力类视频特征,随后采用余弦嵌入损失进行暴力与非暴力类间语义度量,让暴力样本向语义中心聚拢、非暴力样本与该中心离散,缩小数据类内方差的同时,增大了类间方差.该方法在不增加额外数据标注的情况下挖掘数据内部知识,并联合暴力视频分类损失实现具有语义判别性的多任务学习,有效提高了模型的泛化能力.最后,所提方法在VSD2015[7],Violent Flows[8]和RWF-2000[9]三个公开的暴力视频数据集上进行实验验证.实验结果表明,本文方法在三个公开数据集上均有明显效果,较已有算法分别提升了4.79%,0.81%和1.5%.

2 相关工作

在早期的暴力视频识别工作中,研究人员主要采用手工设计的特征.暴力视频识别中最常用的视觉特征主要包括表观特征与运动特征.表观特征包含了视频场景、目标主体等关键性信息,常用的表观特征描述子有尺度不变特征变换(Scale-Invariant Feature Transform,SIFT)、定向梯度直方图(Histogram of Oriented Gradient,HOG)等.而运动特征包含了目标运动的关键信息,对打斗等动作事件具有非常重要的作用.常用的运动特征描述子包括时空关注点(Space-Time Interest Points,STIP)和改进的密集轨迹(improved Dense Trajectories,iDT)等.文献[8]提出暴力流描述符(Violent Flow,ViF)来检测人群中的暴力.文献[10]提出了改进的运动韦伯局部描述符(Motion Improved Weber Local Descriptor,MoIWLD),并通过字典学习进行打斗类暴力检测.此外,爆炸、尖叫、枪击等音频信息对暴

力视频识别也起到了十分重要的辅助作用. 文献[11]提出了一个两阶段的检测过程,其中音频和视频分类器以共同训练的方式结合,最后在决策层进行分数融合. 然而暴力场景一般比较复杂,手工特征表征能力不足以充分描述暴力语义信息,故此类方法识别性能不佳.

近年来,研究者尝试利用深度神经网络进行暴力视频识别. 在2015年的MediaEval情感任务中,Dai等人[12]使用双流网络和长短时记忆(Long Short-Term Memory,LSTM)网络捕获视频帧和运动信息,同时采用决策层融合获得了暴力预测分数,该方法在当年取得了最佳效果. 文献[13]利用目标检测网络和FlowNet[14]构建了双流暴力检测系统,用于识别和定位视频中的打斗行为. 文献[15]建立了一个基于三维卷积的端到端暴力视频识别框架,该框架利用关键帧将视频切割为序列,避免了均匀采样导致的数据冗余和对运动完整性的破坏. 文献[16]为每个暴力类别构建了单独的视觉特征检测器和听觉特征检测器,并且针对不同类别做出了区分,例如血腥场景不需要时间特征、爆炸场景着重检测音频特征等. 多个类别的分类器以集成学习的方式进行组合,形成最终的暴力检测器. 文献[17]提出了一种骨架点交互学习模块来检测视频中的打斗行为,该模块通过构建局部骨架点之间的特定权重来聚焦潜在的暴力区域,并利用多头注意力机制来提升模型的鲁棒性. 文献[18]利用图卷积网络同时捕获暴力视频中的长距离依赖和局部位置关系,并将RGB和音频两种模态特征直接串联进行融合. 文献[19]对三维卷积网络进行改进,加入了多头自注意力和双向长短时记忆网络,用于捕捉过去和未来的时间信息. 文献[20]从神经网络的顶层和底层提取多级特征,然后提出了一个宽密度残差模块学习多级特征的不同组合形式,并利用LSTM捕捉时间信息. 文献[21]提出了基于注意力的暴力视频多模态特征提取模块,并采用决策分数融合方式判断视频是否暴力. Peixoto等人[22]将暴力分解为血液、火焰、枪击等特定概念,并提取表观、运动和音频多模态信息进行融合,这种将暴力视频分解为若干子概念、分而治之的思想,受到数据集标注成本制约,导致可扩展性不足. 文献[23]提取了多模态时空特征并采用多任务学习优化暴力视频识别模型,通过引入基于监督学习的情感分类任务和视音频一致性任务来学习更具判别性的暴力视频表示. 类似地,文献[24]在VSD 2015数据集中增加人工标注的语义一致性标签,通过约束视音频多模态特征融合的边界形成鲁棒的特征表示,从而实现暴力与非暴力视频分类. 但此类方法通常需要额外的标签信息,无疑增加了人为标注工作.

上述方法大多侧重于视频的时序建模,对虐待、打斗等特定的肢体暴力具有不错的识别效果,然而忽略了暴力内容是复杂多样的,数据分布具有较大的类内方差. 构建多个分类器和基于监督信号的多任务学习进行综合决策固然可行,但在一定程度上增加了模型的参数量和数据标注成本. 因此,本文从构建统一的暴力语义角度出发,根据暴力数据集的分布特性,提出了一种基于语义嵌入学习的多模态暴力视频识别模型. 该方法通过中心损失构建暴力语义中心对暴力视频进行聚类,并采用余弦嵌入损失计算暴力与非暴力类的语义度量以增大类间距离,生成具有语义判别属性的视频表示,有效提高了模型的识别精度和泛化能力.

3　本文方法

本文提出了一种基于语义嵌入学习的多模态暴力视频识别模型,如图1所示. 首先,结合暴力视频特点,采用深度神经网络分别提取表观、运动和音频三种模态的特征表示. 随后,设计了多模态特征高效融合模块(Multimodal Efficient Fusion Module,MEFM),通过共享空间映射和多模态特征交互消除模态间异构性,进而获得鲁棒的融合视频表示. 最后,在暴力分类损失的基础上,引入中心损失进行暴力类内语义聚合,缩小类内方差;同时采用余弦嵌入损失进行类间语义度量,进一步增加与非暴力视频的判别距离,通过联合暴力视频判别主任务共同优化,以提高模型的泛化性能.

3.1　多模态特征提取

一般来说,暴力视频中通常包含以下信息:(1)表观信息,主要包括场景信息与人物主体,如血液、枪支、棍棒、有人倒地等;(2)动作信息,其中以打斗、暴乱、虐待等为典型;(3)音频信息,部分暴力事件以音频内容为主,如尖叫、爆炸、碰撞、枪击等. 由于不同模态具有显著的异质性,本文分别采用扩展三维卷积网络(Expand 3D,X3D)[25]、嵌入时序位移模块(Temporal Shift Module,TSM)[26]的二维残差网络、预训练音频神经网络(Pretrained Audio Neural NetworkS,PANNS)[27]提取视频的表观特征 $\boldsymbol{F}_{\mathrm{ap}}$. 运动特征 $\boldsymbol{F}_{\mathrm{mo}}$ 和音频特征 $\boldsymbol{F}_{\mathrm{au}}$,如图1左边部分所示.

3.1.1　表观特征提取

本文采用X3D-L网络提取暴力视频的表观特征. 由于视频帧之间存在大量的冗余信息,通过均匀采样和增加采样序列个数进行平衡. 设视频总帧数为 N,采样得到的视频帧序列的个数为 t,每个视频帧序列长度为 l,第 j 个采样序列的第 i 帧可以用 C_j^i 表示,采样过程为

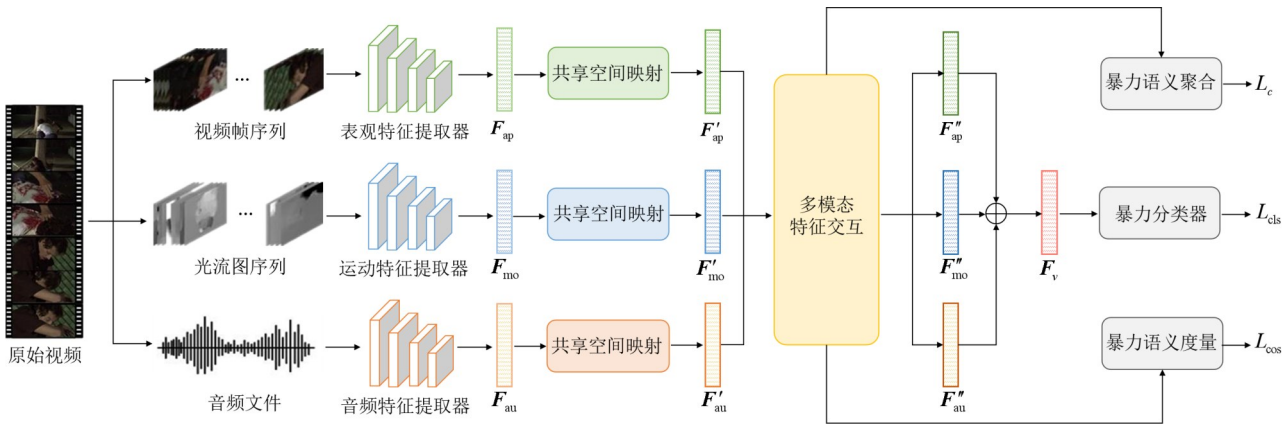


图1 暴力视频识别模型结构图

$$C_j^i = V^{s+(i-1)\times\tau+k}$$
$$s\in[(j-1)\times\frac{N}{t},j\times\frac{N}{t}-l\times\tau],\ k\in[0,\tau] \tag{1}$$

其中，s为采样的起始位置，τ为采样率，k为时域上的随机抖动. 具体来说，首先将整个视频帧均分为t个等长的子区间；其次，在每个子区间上进行均匀采样，获得t个长度为l的片段. 这种方式不仅增加了采样密度，同时保证了暴力帧的覆盖范围，有利于充分提取暴力视频表观特征. 然后，这n个片段被独立地送入X3D-L网络中，得到n个2 048维的特征向量. 最后，本文将n个2 048维的特征向量进行平均得到视频级别的表观特征$\boldsymbol{F}_{\mathrm{ap}}$.

3.1.2 运动特征提取

本文使用光流法提取运动特征. 光流图本身是经过处理的浅层运动特征，其数据量相较于原始视频帧大大减少，使用三维卷积神经网络处理光流图不仅计算量大，还容易产生过拟合现象. 因此，本文采用添加时序位移模块(Temporal Shift Module，TSM)的二维残差网络提取运动特征，在提高运算效率的同时增强模型的时域建模能力. 同时，由于运动特征本身更依赖对整个视频的理解，密集采样不仅计算量大，而且容易出现过拟合，因此，本文采用时域均匀采样处理光流图序列.

首先，采用双流网络[28]中光流图提取方法，分别计算相邻视频帧之间在水平和垂直方向的光流，相对应的水平和垂直两个光流通道被堆叠在一起，形成光流图. 随后，对光流图序列进行均匀采样，具体过程为

$$\widehat{C}^i = \widehat{V}^{s+k}$$
$$s\in[(i-1)\times\frac{N}{l},i\times\frac{N}{l}],\ k\in[0,l] \tag{2}$$

其中，$\widehat{C}^i$表示采样序列中第i个光流图；$\widehat{V}$表示视频的整体光流图序列，长度为N；s表示采样起始位置. 视频的整体光流图序列被平均分为l个子序列，每个子序列中随机抽取一帧，得到长度为l的光流图序列. 最后，将该光流图序列送入添加时序位移模块的ResNet50网络中，得到2 048维的特征向量作为运动特征$\boldsymbol{F}_{\mathrm{mo}}$.

3.1.3 音频特征提取

对于音频信息，本文采用PANNS网络提取音频特征. 首先，使用ffmpeg工具将音频从原始视频中分离出来，并提取该音频的对数梅尔谱图，随后将原始音频和语谱图同时送入PANNS网络中，将二维卷积网络的输出作为音频特征$\boldsymbol{F}_{au}$，其维度为2 048. 具体设置详见第4.2节实验细节.

3.2 多模态高效融合模块

由于多模态特征具有异质鸿沟，采用直接拼接或相加的融合方式可能会破坏原有模态的特征分布，无法充分挖掘不同模态间的互补性. 因此，本文设计了MEFM，如图2所示. 该模块主要包括共享空间映射和多模态特征交互两部分. 前者旨在将不同模态映射至同一语义空间，同时挖掘模态内的多尺度信息；后者旨在捕获多模态特征的互补信息，同时抑制无关的模态间噪声.

3.2.1 共享空间映射

共享空间映射通过对不同模态特征进行多尺度映射，学习得到跨模态的统一表征以抑制不同模态的异质性对融合特征产生的不利影响. 具体计算过程为

$$\begin{aligned}\boldsymbol{F}'_{\mathrm{ap}}&=\delta(f_a(\boldsymbol{F}_{\mathrm{ap}}))+\delta(g_a(\boldsymbol{F}_{\mathrm{ap}}))\\ \boldsymbol{F}'_{\mathrm{mo}}&=\delta(f_m(\boldsymbol{F}_{\mathrm{mo}}))+\delta(g_m(\boldsymbol{F}_{\mathrm{mo}}))\\ \boldsymbol{F}'_{\mathrm{au}}&=\delta(f_u(\boldsymbol{F}_{\mathrm{au}}))+\delta(g_u(\boldsymbol{F}_{\mathrm{au}}))\end{aligned} \tag{3}$$

其中，$\boldsymbol{F}_{\mathrm{ap}}$，$\boldsymbol{F}_{\mathrm{mo}}$，$\boldsymbol{F}_{\mathrm{au}}$分别为表观特征、运动特征以及音频特征；$f(\cdot)$为全连接层；$g(\cdot)$为一维分组卷积层；$\delta(\cdot)$为ReLU激活函数. 由于全连接层具有全局感受野，在线性映射过程中能够捕获模态内的整体信息，而分组卷积沿特征通道进行分组聚合，能够捕获通道内的局部

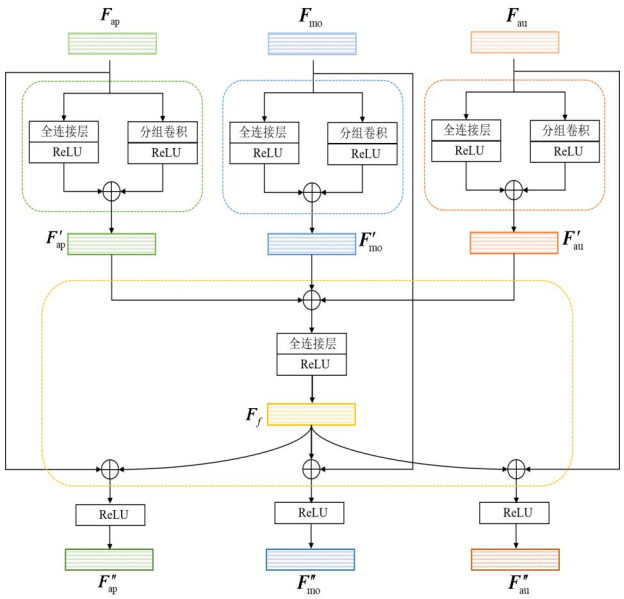


图2　MEFM模块结构示意图

特性. 同时,二者以并行结构耦合,这种设计增加了网络宽度,有利于消除网络训练过程中的过拟合现象. 不同模态的分别经过共享空间映射后,在一定程度上消除了异构性带来的干扰,为下一步多模态特征交互提供了基础.

3.2.2　多模态特征交互

多模态特征交互旨在对不同模态信息进行充分交互,使不同模态学习彼此的互补信息,从而构建更加鲁棒的特征表示. 如何充分利用不同模态之间的信息,而又不对单模态特征本身产生干扰,是本阶段的研究重点. 多模态交互模块结合了残差网络的思想而提出,交互过程可以用式(4)表示:

$$\begin{aligned}\boldsymbol{F}_f&=\delta(h(\boldsymbol{F}'_{\mathrm{ap}}+\boldsymbol{F}'_{\mathrm{mo}}+\boldsymbol{F}'_{\mathrm{au}}))\\ \boldsymbol{F}''_{\mathrm{ap}}&=\delta(\boldsymbol{F}_{\mathrm{ap}}+\boldsymbol{F}_f)\\ \boldsymbol{F}''_{\mathrm{mo}}&=\delta(\boldsymbol{F}_{\mathrm{mo}}+\boldsymbol{F}_f)\\ \boldsymbol{F}''_{au}&=\delta(\boldsymbol{F}_{au}+\boldsymbol{F}_f)\end{aligned}\tag{4}$$

其中,$h(\cdot)$为全连接层;$\delta(\cdot)$为ReLU激活函数. 在共享空间将不同模态的特征进行逐元素相加获得初步融合的特征表示,然后经过全连接层升维至原始特征维度. 该过程使网络进一步挖掘融合特征中的互补信息,以获得多模态的通用表示$\boldsymbol{F}_f$. 随后,利用残差连接将该通用表示分别与原始模态结合,经过ReLU函数激活后,获得增强后的表观特征$\boldsymbol{F}''_{\mathrm{ap}}$、运动特征$\boldsymbol{F}''_{\mathrm{mo}}$和音频特征$\boldsymbol{F}''_{\mathrm{au}}$. 最后,本文将三种模态特征相加得到统一的视频特征表示$\boldsymbol{F}_v$,如式(5)所示:

$$\boldsymbol{F}_v=\boldsymbol{F}''_{\mathrm{ap}}+\boldsymbol{F}''_{\mathrm{mo}}+\boldsymbol{F}''_{\mathrm{au}}\tag{5}$$

3.3　语义嵌入学习

根据采集来源的不同,暴力数据集一般可分为影视作品中的暴力内容和监控镜头下的暴力场景. 前者大多包含场景切换和机位转换等艺术表现手法,暴力的表现形式复杂多变. 而后者通常由固定镜头拍摄,不涉及机位运动和视角变换,以远景和全景画面为主. 由于暴力数据集来源广泛,暴力样本存在类内方差大、类间方差小这一特性,即同为暴力事件,血腥和打斗、枪击和爆炸等存在明显的特征差异. 这一现象对构建通用的暴力识别模型提出了一定挑战. 因此,本文提出了一种基于语义嵌入学习的暴力视频识别方法,在仅利用数据集自身暴力/非暴力标签、未引入其他额外标注信息的前提下,通过充分挖掘数据内部知识实现融入暴力语义的多任务学习.

图3给出了语义嵌入学习方法示意图. 首先,本文通过暴力类内语义聚合构建暴力语义中心,以此聚合暴力类视频特征,实现全局暴力语义的校准和统一. 随后,通过度量暴力/非暴力类间语义,使暴力样本向暴力语义中心聚拢,而非暴力样本与该中心离散,从而形成紧凑的暴力类簇,并与非暴力样本保持语义边距,学习到具有语义判别性的视频特征.

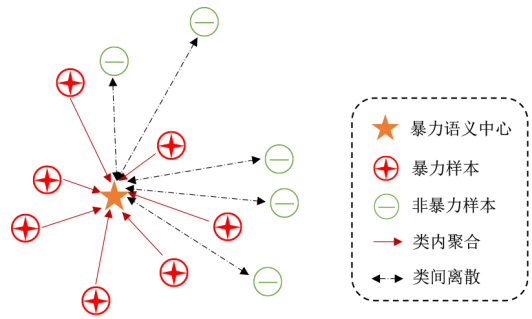


图3　语义嵌入学习方法示意图

在训练阶段,该方法通过将暴力类内语义聚合、类间语义度量这两个任务与暴力视频分类主任务联合学习,降低单个暴力视频分类任务的过拟合风险,从而提升暴力视频识别模型的泛化能力. 在推理阶段,测试视频仅需执行暴力视频分类单个任务即可实现暴力/非暴力类判别.

3.3.1　暴力类内语义聚合

暴力类内语义聚合旨在利用暴力数据的特征分布生成暴力语义中心,该中心被视为暴力类的表示原型,能够充分表征当前数据中的暴力语义. 具体来说,本文使用参数随机初始化中心向量$\boldsymbol{C}$,并采用式(6)更新该暴力语义中心[29],即

$$\begin{aligned}L_c&=\frac{1}{N_v}\sum_{i=1}^{N_v}(\boldsymbol{F}_f^i-\boldsymbol{C})^2\\ \boldsymbol{C}^{t+1}&=\boldsymbol{C}^t-\gamma\cdot\Delta\boldsymbol{C}\end{aligned}\tag{6}$$

其中,$\boldsymbol{F}_f^i$为i个暴力视频特征;N_v为暴力视频的数量;$\Delta\boldsymbol{C}$为该中心向量的参数梯度;γ为该中心向量更新的

步长. 特别地,本文选择MEFM模块中输出的融合特征$\boldsymbol{F}_f$而非整体视频表示$\boldsymbol{F}_v$来构建暴力语义中心,这是由于后者包含了多模态互补信息,某种程度上增加了类内方差且不利于类内语义聚合任务. 最小化中心损失L_c,可以逐步更新该特征空间中的暴力语义中心.

3.3.2 类间语义度量

在获得暴力语义中心$\boldsymbol{C}$后,本文利用该中心动态调整视频样本的特征分布,进一步聚合暴力样本形成密集簇,同时与非暴力样本保持语义边距,这一过程称为类间语义度量. 本文采用余弦距离来实现这一约束,该目标函数如式(7)所示:

$$L_{\cos}=\begin{cases}1-d(\boldsymbol{F}_f,\boldsymbol{C}), & \text{if} \quad y_{\text{cor}}=1\\ \max(0,d(\boldsymbol{F}_f,\boldsymbol{C})), & \text{if} \quad y_{\text{cor}}=-1\end{cases} \tag{7}$$

其中,$d(\cdot)$表示余弦距离度量;$y_{\text{cor}}=1$表示当前样本与中心向量语义一致(即为暴力),$y_{\text{cor}}=-1$表示当前特征与中心向量语义不一致(即为非暴力). 该损失旨在缩小暴力样本与语义中心的差距,同时增大非暴力样本与语义中心的距离,形成具有语义判别属性的特征表示.

3.3.3 暴力视频分类

经MEFM模块输出的统一视频特征表示$\boldsymbol{F}_v$被送入一个简单的全连接层进行暴力分类,获得预测的暴力分数. 这里采用交叉熵来构建暴力分类损失L_{cls},计算过程如式(8)所示:

$$L_{\text{cls}}=-\sum_{i=1}^{B}y_i\log\hat{y}_i \tag{8}$$

其中,B为样本总数;$y_i\in\{0,1\}$为第i个样本的真值标签;$\hat{y}_i$表示该样本的预测分数. 最后,将两种语义嵌入损失与暴力分类损失以多任务学习的形式结合,本文方法的整体目标函数如式(9)所示:

$$L=L_{\text{cls}}+\alpha L_c+\beta L_{\cos} \tag{9}$$

其中,α和β是两个超参数,用于调整中心损失和余弦嵌入损失的权重. 通过优化该整体目标函数,本文模型能在暴力识别的过程中兼顾类别属性和语义特性,使MEFM模块能够动态捕获多模态之间的语义共性,生成具有语义判别属性的融合特征表示,从而提高模型的识别精度和泛化能力.

4 实验结果与分析

为了验证上述模型的有效性与合理性,本文在VSD2015[7],Violent Flows[8]和RWF-2000[9]三个暴力视频数据集上进行了一系列实验与分析. 第4.1节首先介绍了实验所用的数据集及相关评价指标. 第4.2节阐述了实验细节与超参数设置. 第4.3节详细展示了实验结果,包括不同模态的性能比较、多模态特征融合消融实验和语义嵌入多任务学习结果比较. 第4.4节将本文的暴力视频识别模型与其他暴力视频识别算法进行了性能对比,验证所提方法的先进性. 第4.5节对测试结果进行了可视化展示与分析,进一步验证了所提方法的有效性.

4.1 数据集及评价指标

为验证所提方法的有效性,我们在VSD2015[7],Violent Flows[8]和RWF-2000[9]三个具有挑战的公开数据集上进行实验,如表1所示. VSD2015采集自多部好莱坞影视作品和YouTube视频,Violent Flows来自YouTube视频,场景面向体育赛事和游行等群体暴力,RWF-2000主要来源于真实场景中的监控镜头. 不同场景来源的数据集也为验证本文模型的通用性提供了有力支撑.

表1 暴力视频数据集类别分布

数据集	暴力类	非暴力类
VSD2015	502	10 398
Violent Flows	123	123
RWF-2000	1 000	1 000

4.1.1 VSD2015数据集

VSD2015数据集由MediaEval2015情感分析挑战赛发布. 该数据集采集自37部好莱坞电影和30支YouTube视频,共有10 900个视频片段,每一个视频片段剪辑时长在8~12 s. 然而,该数据集中暴力与非暴力类别存在严重的不均衡分布. 不平衡的训练集会使模型偏向于将视频判断为非暴力视频,从而导致模型在非暴力类产生过拟合而在暴力类产生欠拟合的现象. 因此,常见的准确率不能客观反映模型的真实性能,本文采用官方指定的平均精确率(Average Precision,AP),即PR曲线下面积,来评估模型性能. 同时,为缓解类别不均衡带来的偏差,对训练集中的暴力视频采用镜像、旋转、裁剪等组合数据增强方式,以平衡正负样本的数量,从而适应深度模型的训练需求. 对于音频数据,本文采用了Mixup[30]和SpecAugment[31]两种方式进行数据增强.

4.1.2 Violent Flows数据集

Violent Flows(Crowd Violence)是一个公开的群体暴力数据集,所有的视频都源自YouTube网站,视频场景类型比较广泛、人物数量较多、背景嘈杂,是一个比较具有挑战性的暴力数据集. 视频数量共有246个,其中暴力视频和非暴力视频各占一半. 该数据集并未明确区分训练和测试数据,因此本文采用五折交叉验证进行实验. 同时,为满足深度神经网络的训练需求,对训练集进行了数据增强,而测试集保持不变. 此外,该数据集暴力与非暴力类别数量均等,故采用分类准确率(ACCuracy,ACC)评估模型性能,计算式如(10)

所示:

$$\mathrm{ACC}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}\times 100\% \tag{10}$$

其中,TP表示判断正确的暴力样本数;TN表示判断正确的非暴力样本数;FP表示判断为暴力的非暴力样本数;FN表示判断为非暴力的暴力样本数.

4.1.3 RWF-2000数据集

RWF-2000数据集来自真实场景中的监控镜头,没有镜头切换、场景调度等艺术表现手法,同时不包括音频信息. 该数据集包含1 600个训练样本和400个测试样本,视频样本的平均时长为5 s. 同样地,为满足深度神经网络的训练需求,对训练集进行了旋转、镜像、缩放等数据增强方式,而测试集保持不变. 该数据集测试样本类别分布均衡,因此该数据集也是采用准确率作为评价指标.

4.2 实验细节与超参数设置

对于表观通路,视频帧采样序列个数t设置为5,输入序列长度l为32,采样率τ设置为5. 经过时域采样后,得到维度为$32\times 3\times H\times W$的视频帧序列,其中,$H$为原始视频高度,$W$为原始视频宽度,3表示视频帧的RGB通路. 时域采样后,本文对视频序列进行空间采样. 选择视频帧两个边中较短边,将其设置为区间[356,446]中的一个随机数,根据原始视频帧的长宽比对较长的一边进行变换. 然后,在变换后的视频帧序列上进行随机裁剪,得到维度为$32\times 3\times 312\times 312$的视频帧序列. 最后,将$t$个序列送入预训练的X3D-L网络提取片段特征,取$t$个片段特征的均值作为表观特征.

对于运动通路,首先使用TV-L1算法提取密集光流图,将步长设置为2,得到$32\times 2\times H\times W$的水平和垂直方向的光流图序列. 该序列经过与视频帧相同的缩放变换后被随机裁剪为$32\times 2\times 224\times 224$的空间尺寸. 最后,该序列被送入一个带有TSM模块的ResNet50网络获得运动特征.

对于音频通路,本文首先利用ffmpeg工具从原始视频中提取音频文件. 随后,利用窗口大小为1 024、窗口跳距为320样本的汉明窗对音频数据进行预处理,经过短时傅里叶变换得到梅尔谱图,其中每一秒的梅尔谱图含有100个音频帧. 将梅尔谱图映射到覆盖50~14 000 Hz的64个Mel箱,得到对数梅尔谱图. 设音频数据的时长为T,则梅尔谱图的维度表示为$T\times 100\times 64$. 最后,原始音频文件与梅尔谱图被同时送入PANNS网络提取音频特征.

对于MEFM模块,共享空间映射阶段的全连接层含有512个节点,一维分组卷积的输出通道设为512,分组数设为输入维度的1/4;多模态特征交互中的全连接层含有2 048个节点.

由于特征提取器不参与训练,我们采用分层学习率,其中MEFM模块与分类器的初始学习率被设置为1×10^{-5},暴力语义中心更新的学习率γ设置为0.1. 二者均采用余弦衰减策略,训练的迭代批次为256,最大轮数设置为50,采用ADAM优化器进行参数更新,分类器的Dropout设为0.2. 多任务损失权重参数α与β具体设置见第4.3.3节.

4.3 消融实验分析

本节展示了在三个数据集上的系列实验,包括单模态性能对比、多模态特征融合对比、多模态特征融合方法对比、语义嵌入多任务学习对比等消融实验,以充分验证本文方法的有效性.

4.3.1 单模态性能对比

不同模态之间的性能如表2所示. 对于VSD2015数据集,其表观特征的性能最优,音频次之,而运动信息效果最差,这是由于该数据集包含的暴力类别复杂多样,如血腥、打斗、枪击、虐待等暴力事件,表观特征作为主要视觉表示,包含视频中大部分的细节信息. 同时,音频模态的测试AP值与表观模态相差无几,这是由于该数据集含有丰富的音频,对尖叫、枪声、爆炸等听觉暴力内容具有明显的辨别力. 而运动信息主要描述大幅度的肢体动作,对静态暴力和听觉暴力的表达能力有限.

对于Violent Flows数据集,表观特征性能明显优于运动特征,而音频信息识别性能最差,这主要与数据集的内容构成有关. 该数据集侧重群体暴力,运动信息杂乱无章,音频信息更为嘈杂,比如其中来自体育赛事的视频,音频信息混杂了背景音乐和场景音频信息,一定程度上影响了音频信息对暴力行为的识别.

对于RWF-2000数据集,其表观特征的性能最好,运动信息次之,且其运动模态仍有较高的准确率. 这是由于该数据集源于监控镜头,画面中的背景通常静止不动. 因此,画面中的人物行为成为暴力识别的关键因素. 而该数据集中的暴力内容大多为打斗等肢体冲突,暴力事件较为单一,因此单模态也取得了不错的准确率.

表2 单模态特征性能实验结果

模态	VSD2015	Violent Flows	RWF-2000
	AP	ACC /%	ACC /%
表观	0.299 1	96.77	88.25
运动	0.237 8	89.43	82.25
音频	0.295 0	74.78	—

4.3.2 多模态融合性能对比

本节首先分析不同模态使用MEFM模块组合融合的效果,如表3所示. 对于VSD2015数据集,在两两组

合的方式中,表观与音频特征融合效果最好,运动与音频组合次之,而表观与运动的融合性能最差,这一结果与单模态实验也相符合. 一方面说明音频在多模态融合中起到了主导作用,同时也表明表观与运动同属视觉模态,两者的异质性要明显弱于同音频的异质性. 在特征融合过程中,异质性较强的两种模态具有丰富的互补信息. 对于Violent Flows数据集,音频含有背景配乐等干扰信息,一定程度影响了音频特征对暴力视频的判断,但三种模态融合的结果仍高于单一模态的识别准确率,也进一步验证了本文特征融合模块的有效性. 由于RWF-2000数据集不含有音频,使用表观和运动的融合特征在测试集上的准确率为90.25%,较单一的表观特征或运动特征均有明显提升,这也验证了本文提出的MEFM融合模块的通用性.

表3 多模态特征融合消融实验结果

模态			VSD2015	Violent Flows	RWF-2000
表观	运动	音频	AP	ACC /%	ACC/%
√	√		0.329 2	96.32	90.25
√		√	0.372 8	94.32	—
	√	√	0.358 8	89.85	—
√	√	√	0.428 1	97.98	—

此外,为进一步验证本文融合模块的优势,表4给出MEFM模块与几种常见的特征融合方法在3个数据集上的性能比较,加粗数据表示最优结果. 在VSD2015和RWF-2000数据集上,简单的特征拼接和相加融合方式均好于单一模态,表明不同模态之间的确存在互补性. 但在Violent Flows数据集上,由于运动和音频信息含有较多噪声,简单的特征拼接和融合方法结果不如单一的表观结果. DMRN(Dual Multimodal Residual Network)方法[32]通过级联两个全连接层的方式来实现共享空间映射而后采取特征相加输出融合特征,该方法虽然在VSD2015和RWF-2000数据集上取得较好实验结果,但是Violent Flows表观运动和音频三种模态存在干扰,基于DMRN方法仍没有达到很好特征融合结果,且级联方式在一定程度上增加了网络参数. MFB(Multi-modal Factorized Bilinear)[33]采用因式分解简化双线性池化计算,MFH(Multi-modal Factorized High-order pooling)[34]将多个MFB模块级联实现多模态特征间的高阶交互,这两种特征融合方法在三个数据集结果均优于特征拼接、特征融合和单个模态性能,然而此类方法参数量过大,计算过程也较复杂.

在MEFM模块中,共享空间映射阶段单独采用全连接层的效果略差于分组卷积,这是由于分组卷积具有通道内的局部感受野,能够捕获细粒度的特征表示,且其参数量要少于全连接层,能有效抑制过拟合的发生. 本文最后采用基于全连接层和分组卷积并联结构的MEFM模块,在三个数据集上较其他融合方法均取得最佳实验结果,尤其在VSD2015数据集上的性能提升明显. 这是由于VSD2015采用三种模态进行特征融合,其中的音频较表观或运动具有较大的异质性和互补性,通过多模态特征交互后产生的视频表示包含了丰富的多模态信息. 在Violent Flows数据集上,MEFM通过共享空间映射和特征交互两个阶段,以较小参数量实现了表观、含有噪声干扰的运动和音频三种模态的有效融合. 在RWF-2000数据集上,本文方法也更好地融合了表观和运动两种同质性的视觉特征.

表4 多模态特征融合对比实验结果

融合方式	VSD2015	Violent Flows	RWF-2000
	AP	ACC/%	ACC/%
特征拼接	0.387 9	96.33	89.50
特征相加	0.393 1	95.53	89.75
DMRN[32]	0.408 6	95.93	90.00
MFB[33]	0.409 1	97.97	90.00
MFH[34]	0.411 9	97.97	89.75
MEFM (仅全连接层)	0.391 1	97.58	89.00
MEFM (仅分组卷积)	0.392 8	97.13	90.25
MEFM	**0.428 1**	**97.98**	**90.25**

综上所述,本文提出的MEFM融合模块通过全连接层与一维分组卷积并联的结构,在降低参数量的同时增加了网络宽度,一定程度抑制了过拟合现象,在VSD2015, Violent Flows和RWF-2000三个不同公开数据集上均取得了有效验证.

4.3.3 语义嵌入学习消融实验

本节验证了语义嵌入学习对于模型优化的作用,实验结果如表5和表6所示. 首先,针对式(9)中损失函数权重超参数α和β进行了多组实验,结果如表5所示,加粗数据表示最优结果. 在VSD2015数据集中,多任务学习损失权重α设置为20、β设置为15时结果最佳;对于Violent Flows数据集,权重α设置20、β设置为50时,结果最好;对于RWF-2000数据集,α设置为10、β设置为1,结果最优.

在表5基础上,得出表6的语义嵌入学习消融实验结果,加粗数据表示最优结果. 首先,在引入中心损失后,模型在VSD2015和RWF-2000上的性能分别提升了0.35%和0.25%,说明该目标函数促进了暴力样本的类内聚合. 而Violent Flows数据集数据规模较小,在只引入中心损失情况下效果不明显. 在此基础上,通过余弦嵌入损失进一步将暴力样本向语义中心靠拢,形成紧密的样本簇,同时与非暴力样本保持语义边距,实现有

表5　多任务损失权重设置的实验结果

VSD2015			Violent Flows			RWF-2000		
α	β	AP	α	β	ACC /%	α	β	ACC/%
0	0	0.428 1	0	0	97.98	0	0	90.25
1	0	0.431 6	1	0	97.98	1	0	90.50
1	1	0.434 1	1	1	97.98	1	0.1	90.50
1	15	0.443 3	1	20	98.37	1	1.0	91.00
1	30	0.441 6	1	50	98.37	1	2.0	90.75
20	1	0.431 5	20	1	97.98	5	1	91.25
20	**15**	**0.445 5**	20	20	97.98	**10**	**1**	**91.25**
20	30	0.443 2	**20**	**50**	**98.78**	20	1	90.25

效的类间离散. 实验结果表明,在引入余弦损失后,VSD2015的测试AP值提升了1.39%,而Violent Flows和RWF-2000的测试准确率也分别提升了0.8%和0.75%,说明了β加权项对暴力和非暴力样本进行类内聚合和类间排斥的重要性. 综上所述,表6表明了基于中心损失的类内聚合和基于余弦损失的类间语义度量两个辅助任务的有效性. 最后,联合暴力视频分类的多任务学习框架在三个数据集取得了最好结果,有力地验证了语义嵌入学习的有效性和通用性.

表6　多模态特征融合消融实验结果

L_{cls}	L_c	L_{cos}	VSD2015	Violent Flows	RWF-2000
			AP	ACC/%	ACC/%
√			0.428 1	97.98	90.25
√	√		0.431 6	97.98	90.5
√	√	√	**0.445 5**	**98.78**	**91.25**

为进一步验证语义嵌入学习的效果,本文以数据规模较大的VSD 2015和RWF-2000数据集为例,对训练阶段的目标函数收敛情况进行了可视化,如图4所示. 三种损失函数在训练阶段基本收敛,说明模型在训练集上完成了较好的拟合. 特别地,中心损失的下降速率明显快于分类损失和余弦损失,这是由于模型在训练过程中采用了分层学习率设置,该中心损失更新的学习率为0.1,远大于MEFM融合模块和分类器的参数学习率. 较大的学习率可以保证模型在训练过程中快速更新暴力语义中心. 当中心损失趋于收敛时,表明该中心具有全局视野的暴力语义,能够较好地表示该数据集中暴力类别的特征原型. 而融合模块和分类器较小的学习率保证了下游网络的平稳更新,此时余弦嵌入损失也以较小的步长逐渐收敛,使融合特征不至于波动过大,因此能够实现暴力内类语义聚合和类间度量和训练目标.

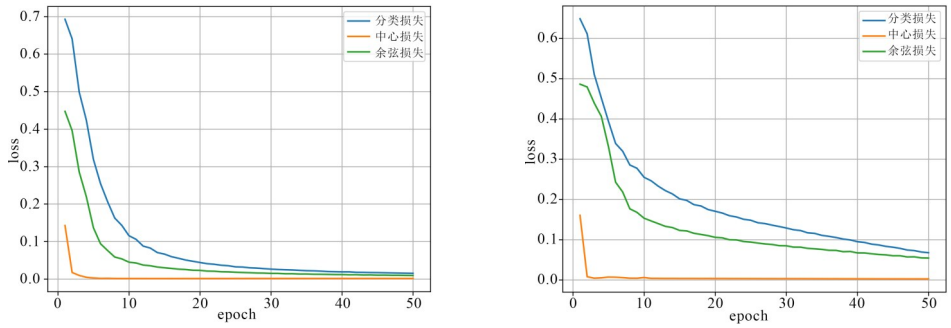
为直观显示语义嵌入的作用,本节还对语义嵌入学习前后的融合特征$\boldsymbol{F}_f$进行了t-SNE可视化,如图5所示. 由于VSD2015测试集中非暴力视频远多于暴力视频,因此本文从非暴力视频从随机选择2 000个样本进行可视化. 在语义嵌入学习前,VSD2015测试集中的暴力样本较为分散,其中部分暴力样本落在了非暴力簇内. 这是由于该数据集中的暴力场景复杂多变,含有较多的困难正样本. 同时,简单的暴力样本由于其较大的类内方差也并未聚合在一起. 在引入语义嵌入学习后,VSD2015中的暴力样本明显地向暴力语义中心聚拢,其中较远处的困难样本也一定程度向聚类中心移动,表明该暴力语义中心促进了具有全局语义的判别特征形成. 而RWF-2000由于其暴力场景的单一性,在语义嵌入学习前已经表现出一定的聚类分布,暴力与非暴力类别的决策边界也较为清晰,因此,语义嵌入学习后的特征分布并未发生明显变化. 图4(b)中心损失的初始值和下降趋势表明,RWF-2000数据集的暴力语义中心并不难习得,即暴力样本的暴力语义较为统一.

4.3.4　本文模型与其他模型对比

为验证所提方法的先进性,本文在VSD2015,Violent Flows和RWF-2000数据集上与已有方法进行了定量比较,结果如表7~9所示,加粗数据表示最优结果.

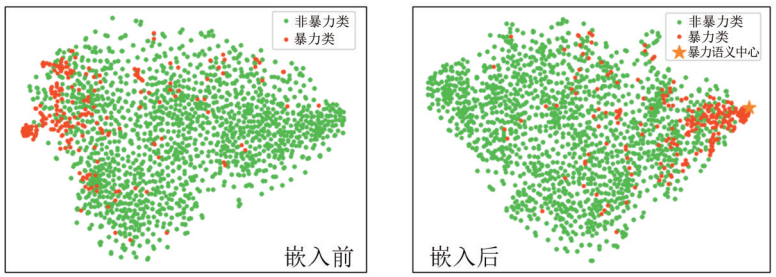
表7表明本文基于语义嵌入学习的多模态暴力识别模型在VSD2015数据集上取得大幅提升,较已有方法[5]提升了4.79%. 这一方面得益于本文采用了合理的多模态特征融合策略,充分挖掘了模态间的互补信息;另一方面,针对暴力视频数量较少的情况,语义嵌入学习在不引入额外标注的情况下,通过多任务学习捕获数据内部的全局暴力语义,一定程度上降低了模型过拟合的现象,提高了模型的泛化能力. 虽然本文模型在该数据集上较已有方法取得了更好的性能,但是由于VSD2015数据集存在严重类别不均衡、暴力类视频较少且数据复杂多样(图5(a)也验证了这种情况)的情况,该数据集的AP值仍整体偏低.

表8列出了在Violent Flows数据集上,本文方法与已有方法的对比结果. Violent Flows数据集是来自YouTube的群体暴力视频,群体打斗运动杂乱无章,音频中也含有背景音乐干扰,虽然利用表观特征就已经可以取得不错的性能,但是如何有效融合不同的多模态信

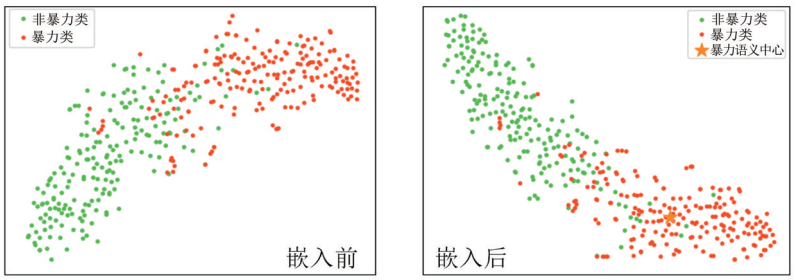


(*a*) VSD2015 (*b*) RWF-2000

图4 多任务学习目标函数收敛情况



(*a*) VSD2015语义嵌入学习前后示意图



(*b*) RWF-2000语义嵌入学习前后示意图

图5 语义嵌入学习特征可视化

表7 本文方法与其他方法在VSD2015数据集的实验结果

方法	AP
Fudan-Huawei[12]	0.295 9
Peixoto 等[22]	0.301 0
文献[24]	0.315 4
TSF_MTL[23]	0.324 2
文献[5]	0.397 6
本文方法	**0.445 5**

息,已成为制约该数据集性能提升的瓶颈;而该数据集规模较小、避免训练模型的过拟合也是要考虑的问题.本文采用的MEFM融合方法在捕获多模态特征的互补信息的同时,抑制模态间无关的噪声干扰.此外,基于语义嵌入的中心损失和余弦嵌入损失发挥了正则化的作用,降低了模型过拟合风险,这使所提方法在该数据集上达到已知的最好结果.

表9列出了本文方法在RWF-2000数据集上的比较结果.与已有研究相比,本文方法在RWF-2000数据

表8 本文方法与其他方法在Violent Flows数据集的实验结果

方法	ACC/%
FGN[9]	88.87
SPIL[17]	94.50
Violence_convLSTM[3]	94.57
ViolenceNet[19]	96.90
文献[20]	97.10
文献[5]	97.97
本文方法	**98.78**

集上依然取得了具有竞争力的结果. 这也表明本文提出的暴力识别模型不仅适用于不同模态的特征融合,同时提出的语义嵌入方法可以自适应地学习到数据集的特征空间分布,并进行有效语义聚合,在数据规模有限的情况改善模型的优化求解,进一步验证了本文所提方法能应对互联网和监控场景等不同素材来源的暴力视频数据集,具有一定通用性和较强的泛化能力.

表9　本文方法与其他方法在RWF-2000数据集的实验结果

方法	ACC/%
文献[21]	85.25
FGN[9]	87.25
VD-Net[35]	88.2
SPIL[17]	89.3
SepConvLSTM[6]	89.75
本文方法	**91.25**

4.4　可视化分析

最后,我们从VSD2015和RWF-2000测试集中输出部分纠正样本进行可视化分析,如图6所示. 其中,VSD2015数据集中的暴力场景图6(*a*)~(*d*)相对复杂,其中部分影片由于艺术表达需要,画面亮度较低,难以捕捉人物运动关系. 图6(*b*)中的枪击由于持续时间较短,且周围环境相对较暗,在未引入语义嵌入学习时,被误判为非暴力视频. 而图6(*c*)和图6(*d*)中的视频是非暴力的,但从画面直接观察难以确定其属于非暴力事件. 图6(*c*)中的双手扼住了一座假人头雕像,尽管该动作倾向于施暴,然而并不会对该雕像造成实质伤害;而图6(*d*)中的拥抱行为极易与打斗等肢体冲突造成混淆. 经过语义嵌入学习后,模型能够挖掘暴力的通用表征,通过从不同暴力类别中捕获语义共性,习得更加抽象的暴力语义,进一步校准了以上误判样本的特征表达,从而实现了错误样本的正确分类.

不同于VSD2015数据集,RWF-2000数据集中的视频为固定机位监控镜头. 相比影视作品中的特写、中近景镜头,此类监控视频大多以俯视全景的角度拍摄,人物主体占画面比例十分有限,因此难以判断暴力事件是否发生. 图6(*e*)和图6(*f*)中发生的打斗和人群暴力由于距离摄像头较远,人物动作的可判别性较低. 语义嵌入学习通过从其他正确判断样本中学习类似场景,进一步动态调整模型对于当前特征的关注区域,从而纠正了错误判别.

(*a*) VSD2015 ACCEDE01398

(*b*) VSD201 -ACCEDE04344

(*c*) VSD2015 MEDIAEVAL00217

(*d*) VSD2015 MEDIAEVAL00323

(*e*) RWF-2000 001809

(*f*) RWF-2000 001852

图6　部分校正后测试样本示意图

5　结论

本文提出了一种基于语义嵌入学习的多模态暴力视频识别模型,有效提升了复杂场景下的暴力识别精度. 在提取表观、运动、音频多模态特征的基础上,本文设计了一种基于共享空间映射和多模态特征交互的特征融合模块MEFM,用于获得鲁棒的多模态融合视频表示. 同时,本文提出一种语义嵌入学习的多任务学习方法,通过中心损失构建全局暴力语义中心;在此基础上,进一步引入余弦嵌入损失聚合暴力样本,缩小暴力类内方差,同时增大与非暴力样本的类间距离. 本文通过结合暴力分类损失,以多任务学习方法优化网络模型,实现了网络的正则化,提高了模型的泛化能力. 实验结果表明,与已有方法相比,本文模型在三个公开暴力视频数据集上均获得性能的提升,为暴力视频识别中的多模态特征融合和内部语义挖掘提供了技术借鉴. 后续研究将考虑构建暴力语义图谱,借助外部知识进一步提升暴力视频识别性能.

参考文献

[1] 闻佳, 王宏君, 邓佳, 等. 基于深度学习的异常事件检测[J]. 电子学报, 2020, 48(2): 308-313.

WEN J, WANG H J, DENG J, et al. Abnormal event detection based on deep learning[J]. Acta Electronica Sinica, 2020, 48(2): 308-313. (in Chinese)

[2] POUR A K, SENG W C, PALAIAHNAKOTE S, et al. A survey on video content rating: Taxonomy, challenges and open issues[J]. Multimedia Tools and Applications, 2021, 80(16): 24121-24145.

[3] SUDHAKARAN S, LANZ O. Learning to detect violent videos using convolutional long short-term memory[C]// 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). Piscataway: IEEE, 2017: 1-6.

[4] HOU C C, WU X Y, WANG G. End-to-end bloody video recognition by audio-visual feature fusion[C]//Chinese Conference on Pattern Recognition and Computer Vision (PRCV). Cham: Springer, 2018: 501-510.

[5] 吴晓雨, 顾超男, 王生进. 多模态特征融合与多任务学习的特种视频分类[J]. 光学精密工程, 2020, 28(5): 1177-1186.
WU X Y, GU C N, WANG S J. Special video classification based on multitask learning and multimodal feature fusion[J]. Optics and Precision Engineering, 2020, 28(5): 1177-1186. (in Chinese)

[6] ISLAM Z, RUKONUZZAMAN M, AHMED R, et al. Efficient two-stream network for violence detection using separable convolutional LSTM[C]//2021 International Joint Conference on Neural Networks (IJCNN). Piscataway: IEEE, 2021: 1-8.

[7] SJÖBERG M, BAVEYE Y, WANG H, et al. The MediaEval 2015 affective impact of movies task[C]//Proceedings of the MediaEval 2015 Workshop. Wurzen: CEUR, 2015: 1-3.

[8] HASSNER T, ITCHER Y, KLIPER-GROSS O. Violent flows: Real-time detection of violent crowd behavior[C]// 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE, 2012: 1-6.

[9] CHENG M, CAI K J, LI M. RWF-2000: An open large scale video database for violence detection[C]//2020 25th International Conference on Pattern Recognition (ICPR). Piscataway: IEEE, 2021: 4183-4190.

[10] ZHANG T, JIA W J, HE X J, et al. Discriminative dictionary learning with motion weber local descriptor for violence detection[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 27(3): 696-709.

[11] LIN J, WANG W Q. Weakly-supervised violence detection in movies with audio and video based co-training[C]// Pacific-Rim Conference on Multimedia. Berlin: Springer, 2009: 930-935.

[12] DAI Q, ZHAO R W, WU Z X, et al. Fudan-Huawei at MediaEval 2015: Detecting violent scenes and affective impact in movies with deep learning[C]//Proceedings of the MediaEval 2015 Workshop. Wurzen: CEUR, 2015: 6-10.

[13] XU Q C, SEE J, LIN W Y. Localization guided fight action detection in surveillance videos[C]//2019 IEEE International Conference on Multimedia and Expo (ICME). Piscataway: IEEE, 2019: 568-573.

[14] DOSOVITSKIY A, FISCHER P, ILG E, et al. FlowNet: Learning optical flow with convolutional networks[C]// 2015 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2016: 2758-2766.

[15] SONG W, ZHANG D L, ZHAO X B, et al. A novel violent video detection scheme based on modified 3D convolutional neural networks[J]. IEEE Access, 2019, 7: 39172-39179.

[16] PEIXOTO B, LAVI B, PEREIRA MARTIN J P, et al. Toward subjective violence detection in videos[C]//ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2019: 8276-8280.

[17] SU Y K, LIN G S, ZHU J H, et al. Human interaction learning on 3D skeleton point clouds for video violence recognition[C]//European Conference on Computer Vision. Cham: Springer, 2020: 74-90.

[18] WU P, LIU J, SHI Y J, et al. Not only look, but also listen: Learning multimodal violence detection under weak supervision[C]//European Conference on Computer Vision. Cham: Springer, 2020: 322-339.

[19] RENDÓN-SEGADOR F J, ÁLVAREZ-GARCÍA J A, ENRÍQUEZ F, et al. ViolenceNet: Dense multi-head self-attention with bidirectional convolutional LSTM for detecting violence[J]. Electronics, 2021, 10(13): 1601.

[20] ASAD M, YANG J, HE J, et al. Multi-frame feature-fusion-based model for violence detection[J]. The Visual Computer, 2021, 37(6): 1415-1431.

[21] ADÃO TEIXEIRA M V, AVILA S. What should we pay attention to when classifying violent videos? [C]//Proceedings of the 16th International Conference on Availability, Reliability and Security. New York: ACM, 2021: 1-10.

[22] PEIXOTO B M, LAVI B, DIAS Z, et al. Harnessing high-

level concepts, visual, and auditory features for violence detection in videos[J]. Journal of Visual Communication and Image Representation, 2021, 78: 103174.

[23] ZHENG Z X, ZHONG W, YE L, et al. Violent scene detection of film videos based on multi-task learning of temporal-spatial features[C]//2021 IEEE 4th International Conference on Multimedia Information Processing and Retrieval (MIPR). Piscataway: IEEE, 2021: 360-365.

[24] LOU J, ZUO D C, ZHANG Z, et al. Violence recognition based on auditory-visual fusion of autoencoder mapping [J]. Electronics, 2021, 10(21): 2654.

[25] FEICHTENHOFER C. X3D: expanding architectures for efficient video recognition[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 200-210.

[26] LIN J, GAN C, HAN S. TSM: Temporal shift module for efficient video understanding[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2020: 7082-7092.

[27] KONG Q Q, CAO Y, IQBAL T, et al. PANNs: Large-scale pretrained audio neural networks for audio pattern recognition[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 2880-2894.

[28] LIU X, YANG X D. Multi-stream with deep convolutional neural networks for human action recognition in videos [C]//International Conference on Neural Information Processing. Cham: Springer, 2018: 251-262.

[29] WEN Y D, ZHANG K P, LI Z F, et al. A discriminative feature learning approach for deep face recognition[C]// European Conference on Computer Vision. Cham: Springer, 2016: 499-515.

[30] ZHANG H, CISSE M, DAUPHIN Y N, et al. Mixup: Beyond empirical risk minimization[C]//Proceedings of the 6th International Conference on Learning Representations. Vancouver: ICLR, 2018:1-13.

[31] PARK D S, CHAN W, ZHANG Y, et al. SpecAugment: A simple data augmentation method for automatic speech recognition[C]//Proceedings of the International Speech Communication Association. Graz: ISCA, 2019: 2613-2617.

[32] TIAN Y P, SHI J, LI B C, et al. Audio-visual event localization in unconstrained videos[C]//European Conference on Computer Vision. Cham: Springer, 2018: 252-268.

[33] YU Z, YU J, FAN J P, et al. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering[C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2017: 1839-1848.

[34] YU Z, YU J, XIANG C C, et al. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering[J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(12): 5947-5959.

[35] ULLAH F U M, MUHAMMAD K, HAQ I U, et al. AI-assisted edge vision for violence detection in IoT-based industrial surveillance networks[J]. IEEE Transactions on Industrial Informatics, 2022, 18(8): 5359-5370.

作者简介

吴晓雨　女,1979年生,辽宁盘锦人. 2004年于吉林大学获得硕士学位,2009年于中国科学院自动化研究所获得博士学位. 现为中国传媒大学信息与通信工程学院教授. 主要研究方向为计算机视觉、视频分析与理解. 中国电子学会会员编号:E190130988M.

E-mail: wuxiaoyu@cuc.edu.cn

蒲禹江　男,1997年生,四川南充人. 中国传媒大学信息与通信工程学院硕士研究生. 主要研究方向为动作识别、视频理解、计算机视觉.

E-mail: pyj2020@cuc.edu.cn