doi:10.3969/j.issn.1001-4616.2023.01.015

双交叉注意力自编码器改进视频异常检测

戚小莎[1],曾　静[2],吉根林[2]

(1.南京师范大学数学科学学院,江苏 南京 210023)
(2.南京师范大学计算机与电子信息/人工智能学院,江苏 南京 210023)

[**摘要**]　针对视频中包含的异常事件数量稀少,信息密集的特征容易被遗漏等问题,本文提出一种双交叉注意力自编码器的视频异常事件检测方法. 首先预处理视频集,提取视频帧中表观和运动特征,然后设计双交叉注意力模块融入自编码器中,使特征图在自编码器中能够更好地关联全局特征. 其次将提取后的特征放入各自的自编码器中学习正常行为,使含有正常事件的视频帧能被模型重构,含有异常事件的视频帧则无法被重构. 最后通过检测模型得到各个视频帧的重构误差从而进行异常事件判定. 该方法可以以局部特征关联全局特征的方式有效提高视频异常事件检测的准确率,通过在多个公开数据集中进行实验验证,证明该方法优于其他同类方法.

[**关键词**]　异常检测,自编码器,帧,重构,深度学习,神经网络,特征提取,融合
[**中图分类号**]TP391　[**文献标志码**]A　[**文章编号**]1001-4616(2023)01-0110-10

Improved Video Anomaly Detection with Dual Criss-Cross Attention Auto Encoder

Qi Xiaosha[1],Zeng Jing[2],Ji Genlin[2]

(1.School of Mathematical Sciences,Nanjing Normal University,Nanjing 210023,China)
(2.School of Computer and Electronic Information/Artificial Intelligence,Nanjing Normal University,Nanjing 210023,China)

Abstract:To solve the problems such as sparse quantity of abnormal events contained in video and information-intensive features are easy ommitted,this paper proposes a dual criss-cross attention auto encoder for video abnormal detection. Firstly,we preprocess the video to extract the apparent and motion features in the video,then design the dual criss-cross attention module and incorporate it into auto encoder,in this way,the features can better correlate the global features. Further,we put the extracted features into the respective auto encoders to learn normal behavior,in this way,video frames containing normal events can be reconstructed by the model and those containing abnormal events cannot be reconstructed. Finally,reconstruction errors of each video frame are obtained by the model to determine the abnormal events. This method can effectively improve the accuracy of abnormal event detection by correlating global features with local features,and it is proved to be better than other similar methods through experimental validation in several public datasets.

Key words:anomaly detection,auto encoder,frame,reconstruction,deep learning,neural network,feature extraction,fusion

随着科技快速发展,监控摄像头的应用范围越来越广,相关视频数量日益增加,以往的视频异常检测方法已经不能满足日益增长的社会需求. 因此在计算机视觉这一领域中,如何创新并改进视频异常检测方法这一课题重新焕发生机,吸引了许多学者前去探索. 在现实场景中,大多数事件是否异常需要取决于当时的场景,且异常事件发生概率十分低,因此视频异常检测[1]仍然存在着许多难点. 近几年,为了应对这些难点,学者通常针对某一特定的异常提出相对应的视频异常检测方法并建立模型. 例如,为了辨别行人是否翻越栏杆或跌倒等相同类型的异常行为或动作[2],研究员通常会采用人体轨迹或者动作识别[3-4]的方法去检测视频中的行人异常与否. 然而,在非空旷场景下,交通通常较为拥挤,会发生行人或车辆被遮挡等现象,这给采取上述两种方法的研究者带来了一定的困扰,使他们在特征提取时只能提取到无遮挡

收稿日期:2022-09-03.
基金项目:国家自然科学基金项目(41971343).
通讯作者:吉根林,博士,教授,研究方向:大数据分析与挖掘技术. E-mail:glji@njnu.edu.cn

的特征,无法完整提取被遮挡的特征,因此无法得到完整准确的特征信息,从而导致异常识别准确率的降低. 针对上述方法所存在的不足,本文将背景减除法与前景提取法相结合,提出了帧差法[5]和光流法[6]相结合的特征提取融合算法—帧流法来提取运动特征,并采用方向梯度直方图(histogram of oriented gradient,HOG)[7]来提取视频中的纹理信息以得到表观特征. 采用神经网络的帧流法能够很好地提取相对完整的运动特征,避免由于运动目标移动缓慢所导致的信息遗漏. HOG 则能够较为清晰地显示视频中所包含的多种纹理信息,且该特征提取算法较为简单,易于上手. 采用多种特征提取算法相结合的方式,能够更好地得到完整的视频特征,从而提高视频异常检测模型的准确度.

在现实场景下,大多数事件都为正常事件,只有极少数异常事件. 因此,本文采用跳跃卷积自编码器[8]并仅用正常事件对模型进行训练,通过重构正常事件使得重构得到的图像与原视频帧相差无几,以便在测试时使得模型无法重构异常帧,即异常图像与重构图像不相似. 由于现有的自编码器只能加强视频帧中相邻像素点的特征信息关联度,为了更好地在局部特征中关联全局上下文特征信息,降低时间和空间复杂度,引入了双交叉注意力模块. 该模块能够使视频帧特征图中的每个像素点都能更好地关联到其他像素点的特征信息. 本文主要贡献如下:

(1)为减少运动特征在提取时被遗漏的可能性,采用新的融合特征——帧流特征作为运动特征;

(2)为提高全局特征与局部特征的关联性,引入双交叉注意力机制以捕获长距离上下文依赖特征信息;

(3)提出一种新的视频异常检测方法——双交叉注意力自编码器(dual criss-cross attention based auto encoder,DCAE),使其能够提高视频异常事件的检测率.

1 相关工作

视频异常检测中关键的步骤为特征提取与模型建立. 其中,特征提取是视频异常能否被准确检测到的关键性指标. 最初研究者通常采用手工设计的方式以提取视频帧特征并建立模型. 由于深度学习[9]发展迅速,深度学习方法被广泛地应用到视频异常检测方向,这大大地提高了特征提取的有效性以及异常检测的准确性. 针对视频中非平稳性的问题,文献[10]首先通过时间递归差分网络进行视频帧预测,其中差分网络被用来处理视频数据的非平稳性,其次对视频异常检测进行自回归移动平均估计,并通过在 3 个空中视频数据集和两个标准异常检测视频数据集上得到结果,证明了所提方法的有效性. 文献[11]提出了新的卷积自编码器架构,该网络结构可以将空间与时间分开表示,以达到分别提取时间与空间信息的目的. 同时为了提高对快速移动异常事件的检测性能,引入方差注意力模块以突出大的运动区域. 该架构在多个数据集中实验并证明有效. 为了解决弱监督下的视频异常检测问题,文献[12]设计了一个用于清理标签噪声的图卷积网络,该网络整合了特征相似性与时间一致性两个异常分析的关键特征,并以端到端的方式进行检测,试验结果表明了该网络的优越性. 文献[13]提出了一种骨架预测网络,将图卷积网络与骨骼特征相结合,更好的提高了模型检测能力. 文献[14]将深度学习与传统方法相结合,提出了一种用于视频异常检测的深度概率模型. 该模型将视频异常检测问题转移到了密度估计问题中,能够将视频异常检测作为一种无监督离群点检测任务来解决,用来解决下潜特征空间中的异常. 文献[15]为了能够充分利用表观和运动特征,提出了一种孪生网络. 该网络能够同时捕捉外观和动作信息,并通过记忆增强模块使异常样本能够更好地被辨认.

自编码器是目前较为常见的深度学习模型之一,主要由编码器和解码器两部分组成. 在训练阶段用不含异常事件的正常视频集通过自编码技术提取全局特征,建立全局高斯模型,通过提取正常视频相邻数帧的结构相似性,建立局部高斯模型. 在测试阶段,将测试视频集分别输入两个高斯模型中,通过马氏距离计算测试视频集与正常视频集的相关性. 在最后的决策阶段,综合两个模型的结果,将两个模型都判定为异常的视频帧判定为异常. 文献[16]结合空间流与时间流提出双流时空自编码器以提取空间-时间特征来检测异常情况,该模型有着较高的精确度. 为获得精确的时空特征,文献[17]提出对抗三维卷积自动编码器来学习正常的时空特征,将事件与视频中学习到的正常模式进行对比,若与正常模式相反则判定为异常事件. 编码器捕捉到视频的空间和时间维度之间的低级关联,并产生代表视觉时空信息的独特特征,解码器从编码后的特征中重新构建原始视频,并以无监督的方式学习正常的时空模式,最终提高自编码器

的重构能力用以辨别异常事件. 文献[18]为了降低视频异常检测方向的计算成本,通过分析 Top-Heavey 设计的便利性,提炼并联合时空训练,比较两种不同的自编码器训练过程,证明使用较小的自编码器网络架构可以较好地减小计算成本. 文献[19]在考虑正常样本多样性的前提下,提出一种时序多尺度自编码器网络,该网络能够建立视频连续帧之间的关联,在保证实时性的同时提升了检测精度.

注意力机制[20]模仿了生物观察行为的内部过程,即一种将内部经验和外部感觉对齐从而增加部分区域的观察精细度的机制. 注意力机制可以快速提取稀疏数据的重要特征,因此被广泛用于计算机视觉任务. 自注意力机制则是对注意力机制的改进,其减少了对外部信息的依赖,更擅长捕捉特征的内部相关性. 为了解决在弱监督下难以在训练时准确识别正常与异常事件的问题,文献[21]提出了能够在特征层面和分数层面将异常实例与正常实例进行区分的相似度注意力网络框架. 该框架将局部时空的不相似性考虑在内,使得它能够在实时场景中检测异常,而不需要额外的窗口缓冲时间. 文献[22]提出了一种未来帧预测的视频异常检测方法,该方法使用生成对抗网络和注意力机制. 其中生成对抗网络中的生成器由 U 型神经网络以及注意力模块演变而来,判别器则由带有自注意机制的马尔科夫模型构成,它可以影响生成器的预测能力从而提高未来视频帧的生成质量. 实验结果表明注意力模块的应用层次越深,检测效果越好. 文献[23]为了能同时描述视频中表观和运动信息,提出了利用注意力机制的多示例学习视频异常检测算法. 该算法利用三维特征 C3D 和光流特征图,通过注意力机制获取特征的权重参数,通过改进的 MIL 排序算法,最终提升了视频异常事件检测的准确度.

2　视频异常检测方法

2.1　处理流程

为解决视频中重要特征遗漏导致视频异常检测准确率下降等问题,本文结合运动特征和表观特征,提取得到较完整的视频特征,其中运动特征通过帧流法提取得到,表观特征则由 HOG 特征表示. 同时,将跳跃连接部分和双交叉自注意力模块[24]融入卷积自编码器中,作为视频异常检测模型的重要组成部分,以此提高局部特征的整体关联性,并降低视频帧的平均重构误差,最终达到提高视频异常检测准确率的效果. 本文方法主要步骤如下(如图 1 所示):

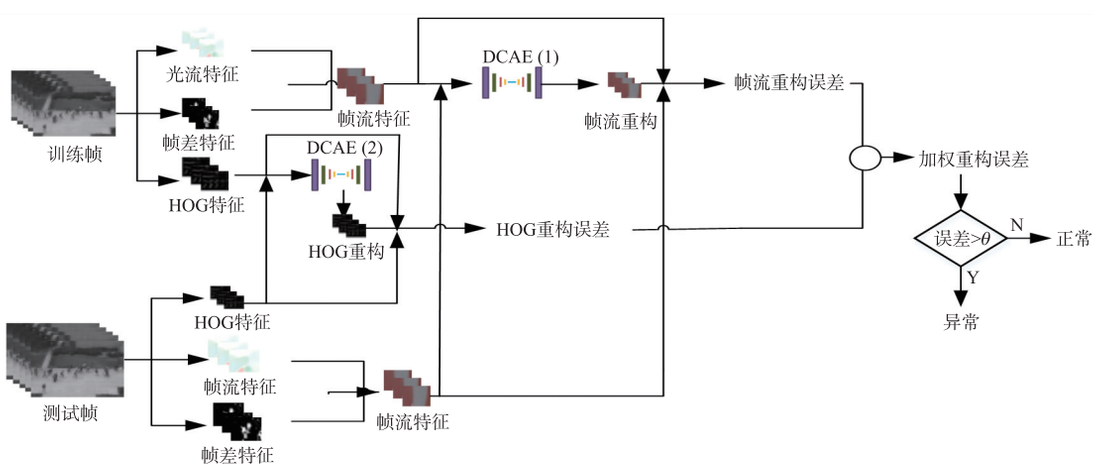


图 1　视频异常检测方法结构

Fig. 1　Structure of video anomaly detection method

(1)数据预处理:首先清洗数据,将原始视频集按不同的视频样本拆分多个帧级别的序列$\{frame_1, frame_2, \cdots, frame_n\}$;

(2)运动特征提取:其次提取视频帧的运动特征. 由于单一的光流特征经常会有信息遗漏等问题,本文采用帧差法,将得到的图像序列中相邻帧对应像素值相减后得到差分图像,并将其二值化,从而有效地得到运动目标位置. 同理,将相邻视频帧输入 Flownet 2 中得到视频帧的光流特征. 将帧差特征与光流特征相结合,则得到本文所需帧流特征,该特征能较好地避免运动物体速度缓慢所导致的信息遗漏;

(3)表观特征提取:再者提取各个视频帧的表观特征. 由于在视频帧中,表观特征为局部目标的表象

和形状,本文采用 HOG 特征提取方法,将待检测的视频帧用梯度或边缘的方向密度分布很好地描述出其表象及形状;

(4)异常检测模型:最后建立视频异常检测模型. 将训练样本提取到的运动特征与表观特征分别输入相对应的自编码器 DCAE 中进行训练,得到训练样本的重构特征及模型. 将测试样本提取到的运动特征与表观特征分别输入已经训练好且相对应的 DCAE 中进行测试. 由于引入了双交叉自注意力模块,DCAE 在训练与测试时能更好地将全局特征与局部特征相关联. 在训练时,DCAE 通过从正常训练样本中提取到的帧流特征以及 HOG 特征学习正常的运动模式,在测试时更准确地重构测试样本特征,得到重构特征,并根据得到的重构误差判定测试样本是否存在异常.

2.2 特征提取

为较完整地提取视频特征,表观特征由 HOG 提取,运动特征则根据帧差特征和光流特征从有限数量的视频卷中提取. 帧差特征和光流特征可以正确地描述运动异常,如人群恐慌、跑步和其他突然变化. 本文将这两个特征融合在一起,得到一个新的融合特征,称为帧流特征. 从图 2 中不难看出,帧流特征可以很好地反映物体的异常运动,降低检测错误率. 光流特征[25]和帧差特征的公式如公式(1)、(2)所示:

图 2 特征融合

Fig. 2 Feature fusion

$$Opt(x,y,t)=Countp^{-1}\sum_{cp=1}^{Countp}\|(\boldsymbol{v}_x^{(cp)},\boldsymbol{v}_y^{(cp)})\|_2, \tag{1}$$

$$Ifd(x,y)=|frame_{fn}(x,y)-frame_{fn-1}(x,y)|. \tag{2}$$

式中,$Countp$ 是视频帧中的像素总数,$\boldsymbol{v}_x^{(cp)}$ 和 $\boldsymbol{v}_y^{(cp)}$ 分别对应于光流的水平和垂直成分. 同样地,f_n 表示帧数.

2.3 双交叉注意力自编码器网络结构

基于双交叉注意力机制和自编码器,本文设计了自编码器 DCAE 网络用于对训练样本特征进行学习,训练好的模型能够检测视频集内是否存在异常事件,若存在,则确定为视频中的哪一帧. DCAE 与传统自编码器的区别在于传统自编码器网络的基本结构为全连接层,这会使得二维图像丢失一定的空间信息;而 DCAE 采用卷积结构对输入特征图进行转换,卷积层能有效地保留所需空间信息,同时在卷积与反卷积中采用跳跃连接,并引入双交叉注意力模块,使得全局特征能够在局部特征中很好地被关联,从而提高视频异常检测模型的准确率. 不同于其他文献采用传统全连接自编码器的方法,本文将训练样本的帧流特征以及 HOG 特征输入相对应的 DCAE 中进行重构,使模型在该过程中学习何为正常事件,以便于在测试时能够更好地重构测试样本的帧流特征以及 HOG 特征并计算其重构误差,这能够更好地判定异常事件是否发生. 自编码器 DCAE 网络结构如图 3 所示,主要由编码器以及解码器构成[26]. 首先,将得到的特征统一压缩为 128×128×3 的图像输入进编码器中. 编码器由两个 3×3 的卷积层以及一个 2×2 的池化层重复构成,每经过一次下采样,特征图的大小变为上一特征图的一半,通道数翻倍,即得到双倍的深度特征图,共

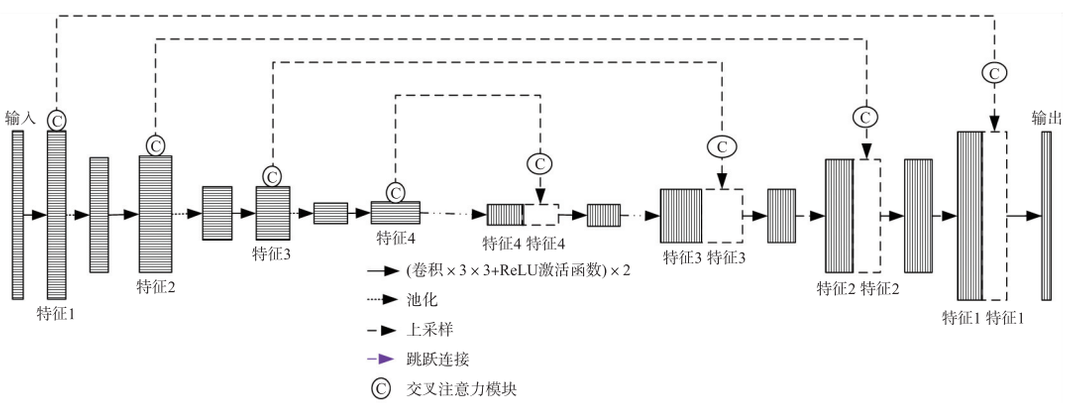


图 3 DCAE 结构

Fig. 3 DCAE structure

进行 4 次迭代以获得不同尺度的深度特征. 编码器中的 3×3 卷积层能够以加权叠加的方式增加输入特征图局部的上下文信息,得到更低维的多个特征向量,最终得到深度特征,该特征相较于输入的特征图更进一步地融合了局部特征的全局观,同时以便于用解码器将其构成新的特征图,并增加理论感受野,使训练模型更加地精确,感受野计算公式如下所示:

$$RF_{k-1}=(RF_k-1)\times s_k+f_k, \tag{3}$$

式中,RF_k 为第 k 层的感受野,f_k 为第 k 层的卷积核大小,$kernel\_size=3$,s_k 为第 k 层的卷积步长,$stride=1$.

其次将得到的深度特征输入双交叉注意力模块中,使得特征图获得更丰富更密集的上下文信息,得到不同尺寸的全局特征与局部特征关联紧密的特征图.

解码器则由一个 2×2 的上采样层以及两个 3×3 的卷积层重复构成,该结构将深度特征重新构建,再采用跳跃连接方式将重构特征与信息密集的特征图连接后输出,得到与输入特征图相同尺寸的图像,从而生成高质量的重构帧.

2.4　交叉注意力

交叉注意力模块在非局部注意力模块的基础上,使用两次注意力加权,并用多个稀疏注意力图代替单个密集注意力图,以减少所需的计算资源.

如图 4 所示,将特征图输入交叉注意力模块后,水平和垂直方向的上下文信息将被收集,分别生成注意力图(上部分)和适配特征图(下部分),并采用加权和的方式将两者进行聚合得到新特征图;将新特征图输入到下一个交叉注意力模块中,该特征图中的每个像素都会从所有其他像素收集信息,以增强像素点的全局关联. 经过两次交叉注意力模块操作后,最终使得每个像素点都关联了特征图中所有像素点的信息. 其中所有交叉注意力模块共享参数以减少额外的参数.

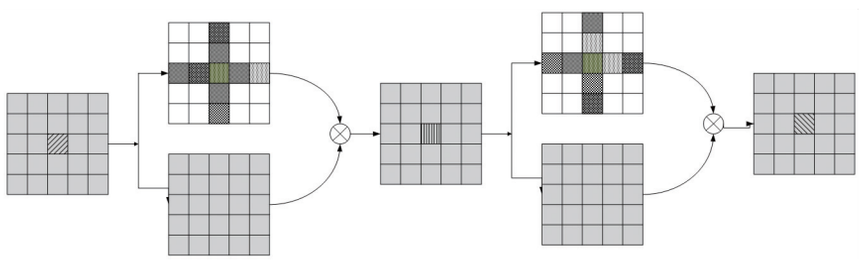


图 4　交叉注意力模块

Fig. 4　Criss-cross attention block

双交叉注意力模块由两个交叉注意力模块构成,交叉注意力模块的内部结构如图 5 所示:

(1)将特征图 x 输入交叉注意力模块后,首先进入两个不同的 1×1 卷积层中进行降维,分别生成两个特征图 h_1,h_2.

(2)得到特征图 h_1,h_2 后,通过 *Affinity* 以及 *Softmax* 运算得到注意力图,如公式(4)、(5)所示:

$$\boldsymbol{d}_{i,u}=\boldsymbol{h}_{1u}\times\boldsymbol{v}_{iu}^{\mathrm{T}} \tag{4}$$

$$softmax(\boldsymbol{z}_a)=\mathrm{e}^{z_a}\Big/\sum_{a=1}^{p}\mathrm{e}^{z_a} \tag{5}$$

式中,$\boldsymbol{d}_{i,u}$表示 $\boldsymbol{h}_{1u}$与 $\boldsymbol{h}_{2iu}$的关联度,$\boldsymbol{h}_{1u}$表示为在特征图 h_1 的空间维度上的每个位置 u 的向量,同理从 h_2 中

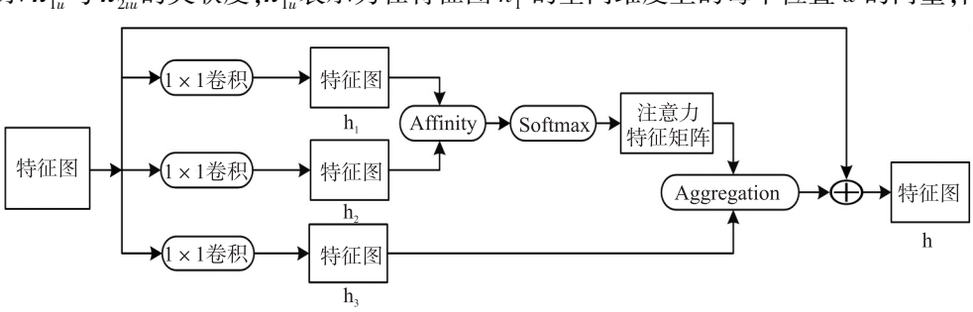


图 5　交叉注意力模块流程

Fig. 5　The process of criss-cross attention block

获得特征向量集合 $\boldsymbol{v}_u$,i 表示为在 $\boldsymbol{v}_u$ 中的第 i 个元素. p 则为类别数,z_a 为输出,a 为输入.

(3)将特征图输入另一个 1×1 卷积层中进行降维生成特征图 h_3,将其与注意力图通过 Aggregation 运算获得具有远处上下文信息的特征图 h,如公式(6)所示:

$$\boldsymbol{h}_u=\sum_i \boldsymbol{A}_{iu}\boldsymbol{y}_{iu}+\boldsymbol{x}_u, \tag{6}$$

$\boldsymbol{A}_{iu}$表示在通道 i 与位置 u 处的标量值,$i\in|\boldsymbol{y}_u|$,$\boldsymbol{y}_u$ 表示特征图 h_3 上的特征向量集合,$\boldsymbol{x}_u$ 表示特征图 x 的空间维度上的每个位置 u 的向量.

(4)最终将特征图与获得远处上下文信息的特征图进行连接,得到关联长距离上下文信息并且每个像素以十字交叉的方式进行信息交互同步的特征图 h.

2.5 损失函数

在所提出的方法中,除了交叉熵损失函数,进一步引入了类别一致性(category consistency)损失函数. 交叉熵损失函数[27]表达式为:

$$L_{ce}=N^{-1}\sum_i\left[(1-z_i)\lg(1-p_i)-z_i\lg(p_i)\right], \tag{7}$$

式中,N 为分类种类数,i 为输入样本,p_i 表示样本 i 预测为正类的概率.

类别一致性损失[28]由拉力损失 L_{var}、推力损失 L_{dis}以及正则化损失 L_{reg}构成,三者函数表达式为:

$$L_{\text{var}}=1/|C|\sum_{c\in C}1/N_c\sum_{i=1}^{N_c}\varphi_{\text{var}}(\boldsymbol{h}_i,\boldsymbol{h}_c), \tag{8}$$

$$L_{\text{dis}}=\left[|C|\times(|C|-1)\right]^{-1}\sum_{c_a}\sum_{c_b}\varphi_{\text{dis}}(\boldsymbol{\mu}_{c_a},\boldsymbol{\mu}_{c_b}), \tag{9}$$

$$L_{\text{reg}}=1/|C|\sum_{c\in C}\|\boldsymbol{\mu}_c\|. \tag{10}$$

其中 C 为输入图像中类的集合,N_c 为 C 中有效元素的数量,$\boldsymbol{h}_i$ 为空间位置 i 的特征向量,$\boldsymbol{\mu}_c$ 为类别 $c\in C$ 的聚类中心,φ 为片段距离函数,φ_{var}与 φ_{dis}的公式如公式(11)、(12)所示:

$$\varphi_{\text{var}}=\begin{cases}\|\boldsymbol{\mu}_c-\boldsymbol{h}_i\|-\vartheta_d+(\vartheta_d-\vartheta_v)^2, & \|\boldsymbol{\mu}_c-\boldsymbol{h}_i\|>\vartheta_d\\ (\|\boldsymbol{\mu}_c-\boldsymbol{h}_i\|-\vartheta_v)^2, & \vartheta_v<\|\boldsymbol{\mu}_c-\boldsymbol{h}_i\|\leqslant\vartheta_d\\ 0, & \|\boldsymbol{\mu}_c-\boldsymbol{h}_i\|\leqslant\vartheta_v\end{cases} \tag{11}$$

$$\varphi_{\text{dis}}=\begin{cases}(2\vartheta_d-\|\boldsymbol{\mu}_{c_a}-\boldsymbol{\mu}_{c_b}\|)^2, & \|\boldsymbol{\mu}_{c_a}-\boldsymbol{\mu}_{c_b}\|\leqslant 2\vartheta_d\\ 0, & \|\boldsymbol{\mu}_{c_a}-\boldsymbol{\mu}_{c_b}\|>2\vartheta_d\end{cases} \tag{12}$$

式中,ϑ_d 与 ϑ_v 代表分界值,设置 $\vartheta_v=0.5$,$\vartheta_d=1.5$.

最终得到类别一致性损失函数如公式(13)所示:

$$L_{cc}=\rho L_{\text{var}}+\tau L_{\text{dis}}+\gamma L_{\text{reg}}, \tag{13}$$

式中,$\rho=\tau=1$,$\gamma=0.001$.

由公式(7)和公式(13)可得单个特征图的 *Loss* 值:

$$L=L_{ce}+L_{cc}, \tag{14}$$

最终可得到输入特征与重构特征的重构误差. 在训练时重构误差越小,重构特征与输入特征越相近,训练后得到的模型准确度就越高.

2.6 视频异常检测方法

在测试阶段,将第 t 帧的光流和 HOG 特征分别放入已训练完成的 DCAE 中计算重构误差. 当重构误差大时,测试样本存在异常事件,反之,则正常. 在得到 HOG 特征重构误差 L_{hog}和帧流重构误差 L_{ff}后,通过公式(15)计算总重构误差:

$$Loss=\alpha L_{hog}+\beta L_{ff},\quad \alpha+\beta=1, \tag{15}$$

由公式(16)可判定视频帧中是否发生异常事件:

$$F=\begin{cases}0, & Loss\leqslant\theta\\ 1, & Loss>\theta\end{cases} \tag{16}$$

式中,θ 为重构误差阈值,若 $Loss>\theta$,该帧发生异常事件,若 $Loss\leqslant\theta$,则表示都为正常事件. 双交叉注意力自

编码器的视频异常检测模型训练过程如算法 1 所示.

算法 1 双交叉注意力自编码器的训练过程

Algorithm 1 Training Process of DCAE

输入:训练视频帧中提取的特征 fi,迭代次数 iter,训练帧数 n;
输出:重构特征 $reconfi$,重构误差 Li
1 for $i=0$ to n:
2 {for $e=0$ to iter:
3 {connection and crop f_i//连接并裁剪
4 $reconfi$=forward($DCAE$, f_i)//得到重构特征
5 $Li=Lce+Lcc$;//计算每帧重构误差
6 $bp(DCAE,reconfi,Li)$;//更新参数
7 return $reconf_i$, Li

3 实验与结果

3.1 实验设置

为了验证自编码器 DCAE 的有效性和准确性,在 Pycharm 中使用 PyTorch 框架和 NVIDIA GeForce GTX3080ti GPU 进行仿真实验,数据集则采用了两个公开数据集(如表 1 所示):CUHK Avenue[29] 和 UCSD Ped2[30]. UCSD 数据集的场景为人行道,Ped2 是行人与摄像头平行移动的场景,场景中的异常事件包括人行道上的自行车和轮椅、行人奔跑、滑板和汽车等. CUHK Avenue 数据集的场景为校园大道,其中包含的异常事件有:行人奔跑、走错方向、卡车、自行车、可疑物品等. 在两个数据集中,训练集都只包含正常事件,测试集则包含正常和异常事件.

表 1 实验数据集

Table 1 Experimental data sets

数据集	场景	尺寸(分辨率)	时间/min
CUHK Avenue	校园大道	640×360	30
UCSD Ped2	人行道	360×240	10

接收者操作特征曲线(receiver operating characteristic curve,ROC)用来衡量 DCAE 网络检测异常的准确率,由真阳性率(true positive rate,TPR)和伪阳性率(false positive rate,FPR)共同构成,如公式(17)、(18)所示:

$$TPR=TP/(TP+FN), \tag{17}$$

$$FPR=FP/(TN+FP). \tag{18}$$

式中,TP 表示真阳性即正确的肯定,TN 表示真阴性即正确的否定,FP 表示假阳性即错误的肯定,FN 表示假阴性即错误的否定.

因此,实验结果使用两种评价标准:(1)曲线下面积(area under the curve,AUC);(2)等错误率(equal error rate,EER). 两者为类似的性能评价指标,当 EER 趋向于 0 时,AUC 趋向于 100%.

3.2 消融实验

为了验证双交叉注意力模块在整个自编码器网络中的作用,检测该模块是否有利于提高视频异常事件的检测准确率,将引入双交叉注意力模块的视频异常检测模型实验结果与未引入该模块的模型实验结果进行对比,如表 2 所示.

表 2 双交叉注意力模块对检测的影响

Table 2 Effect of dual criss-cross attention on detection

	AUC/%	
	Avenue	Ped2
有双交叉注意力	84.93	95.21
无双交叉注意力	83.85	93.77

由表 2 可知,相较于无双交叉注意力模块的视频异常检测模型,引入双交叉注意力模块的模型能够更加准确地检测视频异常事件.

3.3 实验结果

本文将自编码器 DCAE 与当前先进的几种视频异常检测方法进行了对比. 实验结果表明,DCAE 有更高的异常事件检测准确率和更低的误判率,明显优于其他对比方法. 如图 6 所示,标记处为视频发生异常事件区域.

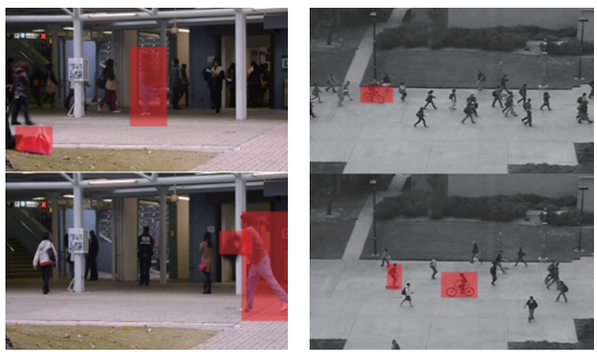
双交叉自注意力自编码器的视频异常检测方法在数据集 CUHK Avenue 下的实验结果如表 3 与图 7

所示. 表 3 可观察到在该数据集下 DCAE 与其他方法各自的 AUC 和 EER 评估结果. 相应的 ROC 曲线则在图 7 中表示.

表 3　CUHK Avenue 下与其他方法比较结果

Table 3　Comparison with other methods for CUHK Avenue　%

检测方法	EER	AUC
AD-ConvAE[31]	27.7	81.65
ConvLSTM[32]	20.7	80.3
sRNN-AE[33]	–	83.48
Sparse Combinations	–	80.9
Proposed	19.9	84.93



CUHK Avenue　　UCSD Ped2

图 6　异常事件可视化

Fig. 6　Visualization of abnormal event

由表 3 可知,本文所提方法在一系列视频异常检测方法中,检测 CUHK Avenue 数据集的准确率与等错误率都有一定的优化.

图 8 为数据集 UCSD Ped2 的 ROC 曲线,并由表 4 对相应的 AUC 和 EER 进行评估. 从其中可知,DCAE 相较于其他的视频异常检测算法更加具有优越性.

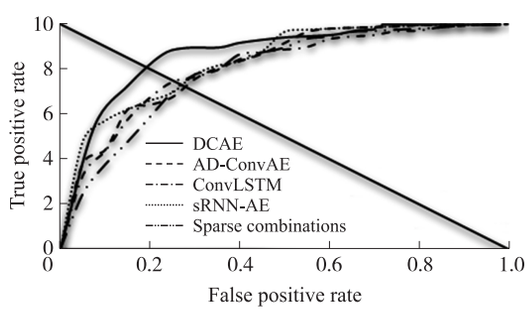


图 7　CUHK Avenue 的 ROC 曲线

Fig. 7　Comparison with other methods for CUHK Avenue

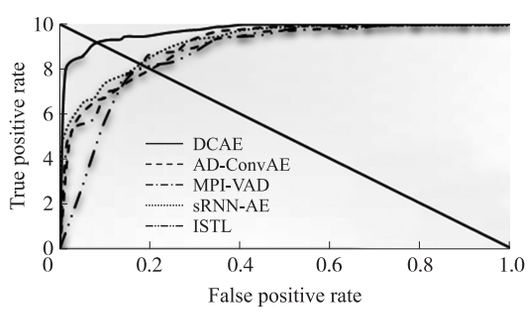


图 8　UCSD Ped2 的 ROC 曲线

Fig. 8　Comparison with other methods for UCSD Ped2

从实验结果中可看出视频中的异常事件有一定的概率会被判定为正常,从而影响检测精确率. 这是因为有些在视频片段中被认为异常的事件,在单视频帧中会被认定为正常事件. DCAE 基于自编码器模型改进而来,其泛化能力过强,输入数据又为单一视频帧,从而导致需要视频片段才能确定的异常事件在单帧中被当成正常事件进行重构,其重构帧与真实帧间的误差低于阈值,以至于被误判为正常事件.

表 4　UCSD Ped2 下与其他方法比较结果

Table 4　Comparison with other methods for UCSD Ped2　%

检测方法	EER	AUC
AD-ConvAE	15.6	90.45
sRNN-AE	–	92.21
ISTL[34]	8.9	91.1
MPI-VAD[35]	16.8	87.5
Proposed	9.8	95.21

4　结论

本文提出了一种基于双交叉注意力自编码器 DCAE 的视频异常事件检测方法. 通过在自编码器中引入双交叉注意力模块,重新构建输入的视频特征,改善了特征图中像素点与全局信息的关联度,提升了自编码器网络的训练效果和检测率. 与近几年其他视频异常检测方法相比,自编码器 DCAE 具有更高的检测准确率以及较低的等错误率. DCAE 在保证精确度的同时,将数据集中无效及无用的信息在训练及测试前提前剔除,并采用连续的稀疏注意力图代替普通的单密集注意力图,这能很好地减少使用的计算资源,从而保证整个方法的运行速度. 由于采用的验证数据集无法包含所有的正常事件以及异常事件,下一步将考虑当特征类似的异常事件出现时,采用自训练模块以提高视频异常检测识别精度.

[参考文献]

[1]　ZAIGHAMZAHEER M,JIN H,LEE S,et al. A brief survey on contemporary methods for anomaly detection in videos[C]//

Proceedings of International Conference on Information And Communication Technology Convergence. Allahabad, India, 2019: 472-473.

[2] CHIMAN D, DINESH K. A review of state-of-the-art techniques for abnormal human activity recognition[J]. Engineering applications of artificial intelligence, 2019, 77: 21-45.

[3] SARAH A, GABRIEL P, SILVIO J, et al. Fight detection in video sequences based on multi-stream convolutional neural networks [C]//Proceedings of 32nd SIBGRAPI Conference on Graphics, Patterns and Images. Rio de Janeiro, Brazil, 2019: 8-15.

[4] ZHU C, WANG Y K, PU D B, et al. Multi-modality video representation for action recognition[J]. Journal on Big Data, 2020, 2(3): 95-104.

[5] XU D M, HAN G G. Application of improved ViBe algorithm in vehicle detection[C]//International Conference on Artificial Intelligence and Pattern Recognition. Beijing, China, 2021: 199-204.

[6] ILG E, MAYER N, SAIKIA T, et al. FlowNet 2.0: Evolution of optical flow estimation with deep networks[C]//Proceedings of Conference on Computer Vision and Pattern Recognition. Hawaii, USA, 2017: 1647-1655.

[7] BUKALA A, KOZIARSKI M, CYGANEK B, et al. A study on pattern recognition with the histograms of oriented gradients in distorted and noisy images[J]. Journal of universal computer science, 2020, 26(4): 454-478.

[8] SU Z B, LI W, MA Z, et al. An improved U-Net method for the semantic segmentation of remote sensing images[J]. Applied intelligence, 2022, 52: 3276-3288.

[9] 徐涛, 田崇阳, 刘才华. 基于深度学习的人群异常行为检测综述[J]. 计算机科学, 2021, 48(9): 125-134.

[10] PILLAI G, SEN D. Anomaly detection in nonstationary videos using time-recursive differencing network-based prediction[J]. IEEE geoscience and remote sensing letters, 2022, 19: 1-5.

[11] CHANG Y P, TU Z G, XIE W, et al. Video anomaly detection with spatio-temporal dissociation[J]. Pattern recognition, 2022, 122: 2-13.

[12] LI N N, ZHONG J X, SHU X J, et al. Weakly-supervised anomaly detection in video surveillance via graph convolutional label noise cleaning[J]. Neurocomputing, 2022, 418: 154-167.

[13] LUO W X, LIU W, GAO S H. Normal graph: spatial temporal graph convolutional networks based prediction network for skeleton based video anomaly detection[J]. Neurocomputing, 2021, 444: 332-337.

[14] OUYANG Y Q, VICTOR S. Video anomaly detection by estimating likelihood of representations[C]//Proceedings of 25th International Conference on Pattern Recognition. Milan, Italy, 2020: 8984-8991.

[15] 李自强, 王正勇, 陈洪刚, 等. 基于外观和动作特征双预测模型的视频异常行为检测[J]. 计算机应用, 2021, 41(10): 2997-3003.

[16] LI T, CHEN X Y, ZHU F S, et al. Two-stream deep spatial-temporal auto-encoder for surveillance video abnormal event detection[J]. Neurocomputing, 2021, 439: 256-270.

[17] SUN C, JIA Y D, SONG H, et al. Adversarial 3D convolutional auto-encoder for abnormal event detection in videos[J]. IEEE transactions on multimedia, 2021, 23: 3292-3305.

[18] ESQUIVEL E, ZAVALETA Z. An examination on autoencoder designs for anomaly detection in video surveillance[J]. IEEE access, 2022, 10: 6208-6217.

[19] 吕浩, 易鹏飞, 刘瑞, 等. 用于视频异常检测的时序多尺度自编码器[J]. 图学学报, 2022, 43(2): 223-229.

[20] CHAUDHARI S, MITHAL V, POLATKAN R, et al. An attentive survey of attention models[J]. ACM transactions on intelligent systems and technology, 2021, 12(53): 1-32.

[21] SNEHASHIAS M, SRIJAN D, FRANCOIS B. DAM: dissimilarity attention module for weakly-supervised video anomaly detection [C]//Proceedings of IEEE International Conference on Advanced Video and Signal Based Surveillance. Washington, USA, 2021: 1-8.

[22] WANG C X, YAO Y X, YAO H. Video anomaly detection method based on future frame prediction and attention mechanism [C]//Proceedings of IEEE 11th Annual Computing and Communication Workshop and Conference. Online, 2021: 405-407.

[23] 魏思倩, 吉根林, 许振, 等. 利用注意力机制的多示例学习视频异常检测[J/OL]. 沈阳市, 小型微型计算机系统, 2021. (2021-10-18)[2022-04-05]. http://kns.cnki.net/kcms/detail/21.1106.tp.20211014.1237.002.html.

[24] FENG S T, ZHUO Z S, PAN D R, et al. CcNet: A cross-connected convolutional network for segmenting retinal vessels using multi-scale features[J]. Neurocomputing, 2020, 392: 603-612.

[25] PIGA N, ONYSHCHUK Y, PASQUALE G, et al. ROFT: real-time optical flow-aided 6D object pose and velocity tracking[J]. IEEE robotics and automation letters, 2022, 7(1): 159-166.

[26] CHEN W Y, PODSTRELENY P, CHENG W H, et al. Code generation from a graphical user interface via attention-based encoder-decoder model[J]. Multimedia systems, 2022, 28: 121-130.

[27] ZHANG C X, HU Y H, ZHU X M. Anomaly detection for user behavior in wireless network based on cross entropy[C]// Proceedings of IEEE 12th International Conference on Ubiquitous Intelligence and Computing and IEEE 12th International Conference on Autonomic and Trusted Computing and IEEE 15th International Conference on Scalable Computing and Communications and Its Associated Workshops. Goyang, South Korea, 2015: 1258-1263.

[28] BERT D, DAVY N, LUC V G. Semantic instance segmentation with a discriminative loss function[EB/OL]. 2017. http://arxiv.org/pdf/1708.02551.

[29] LU C W, SHI J P, JIA J Y. Abnormal event detection at 150 FPS in MATLAB[C]//Proceedings of IEEE International Conference on Computer Vision. Portland, USA, 2013: 2720-2727.

[30] VIJAY M, LI W X, VIRAL B. Anomaly detection in crowded scenes[C]//Proceedings of Computer Vision & Pattern Recognition. San Francisco, USA, 2010: 1975-1981.

[31] 李欣璐,吉根林,赵斌. 基于卷积自编码器分块学习的视频异常事件检测与定位[J]. 数据采集与处理, 2021, 36(3): 489-497.

[32] YONG S, YONG H. Abnormal event detection in videos using spatiotemporal autoencoder[C]//Proceedings of the International Symposium in Neural Networks. Hokkaido, Japan, 2017: 189-196.

[33] LUO W X, LIU W, LIAN D Z, et al. Video anomaly detection with sparse coding inspired deep neural networks[J]. IEEE transactions on pattern analysis and machine intelligence, 2021, 43(3): 1070-1084.

[34] RASHMIKE N, DAMMINDA A, DASWIN D, et al. Spatiotemporal anomaly detection using deep learning for real-time video surveillance[J]. IEEE transactions on industrial informatics, 2020, 16(1): 393-402.

[35] XU Z, ZENG X Q, JI G L. Improved anomaly detection in surveillance videos with multiple probabilistic models inference[J]. Intelligent automation and soft computing, 2022, 31(3): 1703-1717.

[责任编辑:陆炳新]