融合包注意力机制的监控视频异常行为检测

肖进胜 [1] 申梦瑶 [1] 江明俊 [1] 雷俊峰 [1] 包振宇 [1]

摘 要 针对监控视频中行人非正常行走状态的异常现象, 提出了一个端到端的异常行为检测网络, 以视频包为输入, 输出异常得分. 时空编码器提取视频包时空特征后, 利用基于隐向量的注意力机制对包级特征进行加权处理, 最后用包级池化映射出视频包得分. 本文整合了 4 个常用的异常行为检测数据集, 在整合数据集上进行算法测试并与其他异常检测算法进行对比. 多项客观指标结果显示, 本文算法在异常事件检测方面有着显著的优势.

关键词 异常检测, 视频包, 时空特征, 注意力机制

引用格式 肖进胜, 申梦瑶, 江明俊, 雷俊峰, 包振宇. 融合包注意力机制的监控视频异常行为检测. 自动化学报, 2022, **48**(12): 2951−2959

DOI 10.16383/j.aas.c190805

Abnormal Behavior Detection Algorithm With Video-bag Attention Mechanism in Surveillance Video

XIAO Jin-Sheng[1] SHEN Meng-Yao[1] JIANG Ming-Jun[1] LEI Jun-Feng[1] BAO Zhen-Yu[1]

Abstract Aiming at the detection of the abnormal behavior pedestrians in surveillance videos, this paper proposes an end-to-end abnormal behavior detection network. It takes video bags as input, and anomaly score as output. The spatio-temporal encoder is used to extract the features of the video bag, then use the attention mechanism based on the hidden vector to weight the different elements in the bag-level feature, and finally use the bag-level pooling to obtain the video bag anomaly score. Four commonly used anomaly detection datasets are integrated and used to test and compare the performance of different anomaly detection algorithm. The results of multiple objective indicators show that our algorithm has significant advantages in anomaly detection.

Key words Anomaly detection, video-bag, spatiotemporal feature, attention mechanism

Citation Xiao Jin-Sheng, Shen Meng-Yao, Jiang Ming-Jun, Lei Jun-Feng, Bao Zhen-Yu. Abnormal behavior detection algorithm with video-bag attention mechanism in surveillance video. *Acta Automatica Sinica*, 2022, **48**(12): 2951−2959

随着监控摄像头的广泛使用, 其在维护社会安全、进行法律取证等方面凸显日益重要的作用. 由于监控视频与审查人员数量不对等, 大量监控视频无法得到有效处理. 因此对视频监控进行自动智能分析十分必要. 视频在时间维度上蕴含丰富信息, 但其高维度的特性也使特征表达更为复杂而难以完整地提取; 此外, 视频中包含大量的交互行为, 增加了视频处理的难度. 在异常行为的界定上, 类似行为在不同场景会被认为是不同的行为类型. 例如汽车行驶在车行道上属于正常行为, 出现在人行道中则属于异常行为. 异常行为种类多、时间短、判别困难, 很难找到大量数据进行深层网络训练. 这些问题给视频异常行为检测带来巨大的挑战.

本文主要针对步行行人群体场景进行异常行为检测. 异常行为定义为缓慢人群中的快速运动如骑车、奔跑、高空坠物等. 本文将用于异常检测的经典数据集进行整合和事件重标注, 使其与本文所要解决的问题契合, 并在该数据集上进行与其他算法的对比测试. 测试结果表明, 本文算法在指标上有较大的优势. 本文第 1 节介绍了异常事件检测方面的相关工作; 第 2 节介绍本文提出的异常检测算法, 给出算法中各模块和损失函数的具体介绍; 第 3 节首先说明数据集的生成, 然后分析了本文算法的实验效果, 并与其他算法进行对比, 在多个指标上进行分析与评价; 第 4 节给出全文总结.

1 相关工作

视频异常事件检测受多种因素掣肘而极具挑战

收稿日期 2019-11-25 录用日期 2020-03-25
Manuscript received November 25, 2019; accepted March 25, 2020
国家重点研发计划 (2016YFB0502602, 2017YFB1302401)资助
Supported by National Key Research and Development Program of China (2016YFB0502602, 2017YFB1302401)
本文责任编委 桑农
Recommended by Associate Editor SANG Nong
1. 武汉大学电子信息学院 武汉 430072
1. School of Electronic Information, Wuhan University, Wuhan 430072

性, 例如: 异常视频数量少、异常事件难以定义且种类繁多. 因此, 很难找到一个具有概括力的模型囊括所有异常事件, 所以学者们大多从提取视频特征的角度出发, 寻找异常视频最具表现力的特征表示方法. 视频特征的人工提取方法着重提取分析低层次视觉特征, 如导向梯度直方图[1]、光流图[2]、主成分分析[3] 等. 文献 [4] 中探索了多流手动特征进行视频表达的方法, 构造了由时空向量和位置向量组成的视频立体表达结构, 并改进编码方法使其表征能力更强. 但人工特征提取步骤复杂且不全面.

随着深度学习的快速发展, 利用神经网络[5] 进行自动特征提取成为研究热点. 基于视频固有特性, 有学者提出使用 3D 卷积神经网络[6] 实现对视频时空特征的提取. 文献 [7] 学习单目标个体的外形和三维梯度运动信息, 分类时将正常行为模式划分为多个伪异常类, 通过多个一对多二分类器来进行分类. 文献 [8] 提出了一种多尺度时间递归神经网络, 该网络对划分的多网格区域提取的特征进行空间维度和时间维度的关系建模, 以此来进行人群异常事件的检测和定位.

基于重建误差的方式进行异常检测的代表当属自编码网络[9], 通过计算解码器输出数据与编码器输入数据之间的欧氏距离来判断输入偏离正常分布的程度, 以此判断是否异常. 文献 [10] 中设计了一个时空自编码器, 通过空间卷积与基于卷积的长短期记忆 (Convolutional long-short term memory, ConvLSTM)[11] 结构来对视频序列进行编码, 再利用具有对称结构的解码器将视频编码转换成原始图像序列, 通过计算解码图像与原始图像的欧氏距离得到重建误差, 再得到异常分数. 文献 [12]提出了一种变分自编码器 (Variational autoencoder, VAE), 利用视频编码结果拟合分布函数, 对分布参数而非编码值本身进行建模. 文献 [13] 在稀疏去噪自编码网络的基础上, 添加了梯度差约束条件, 使网络的性能在全局异常行为检测方面更有效.

文献 [14] 中将带有异常片段的视频分为若干视频段组成多个视频实例, 设计了全连接网络将 C3D 网络[15] 提取的视频特征映射成异常分数, 根据排名损失使含有异常片段的实例得分高于仅含正常片段实例的得分.

本文模型训练使用了包含异常和正常的视频, 使模型预测更有针对性, 以视频包为网络输入, 通过端到端的网络映射出视频包得分, 输入形式灵活. 本文设计的端到端异常检测网络算法具有以下几点贡献:

1) 以时空自编码网络为基础、视频包为视频提取的基本单位, 以提取更多有效特征, 同时解决了时空自编码器不能实时输出视频异常得分的问题, 扩展了应用场景.

2) 引入多实例排名框架中重要节点得分高的思想, 利用基于隐向量的注意力机制处理视频包时空特征, 以突出重要特征, 弱化无关信息对检测结果的干扰.

3) 采用交叉熵损失耦合铰链损失函数, 优化网络模型预测效果.

2 本文算法

本文设计了一个端到端的异常检测网络, 如图 1 所示, 以视频包为输入. 视频包由视频中顺序排列的连续 τ 帧图像组成, 在时间维度上以步长 1 滑动窗口, 可得到新的视频包. 利用视频包进行特征提取可以在进行空间纹理特征分析的同时保留时间信息, 使提取的特征更全面、更有表现力.

对于视频包, 首先提取单帧空间特征和多帧时间特征. 为了更好地实现特征融合, 利用三维卷积核在时间和空间维度再同时作加权特征融合, 从而得到整个视频包的时空特征表示. 之后采用加入 Dropout 操作的全连接层将特征转换到一个更有表

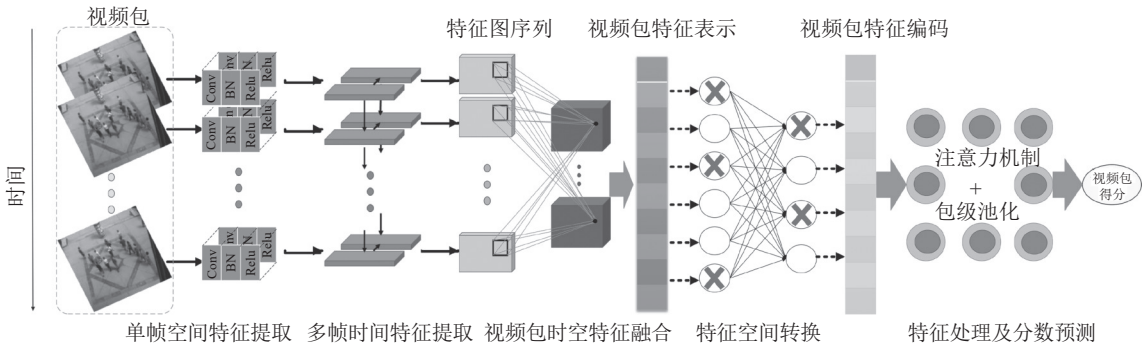


图 1 异常行为检测网络架构

Fig. 1 The framework for abnormal behavior detection

现力的空间中, 同时降低模型过拟合风险; 对于视频包的新特征表示, 根据其自身特点利用注意力机制对其各元素进行加权处理, 再利用包级池化操作将视频包级别的特征映射为该视频包的异常分数 B. 具体操作为

$$B = Pool(K_{\text{atten}}(F_{\text{bag}}(P_{\text{pre}}(I)))) \tag{1}$$

其中, I 为输入图像, P_{pre} 进行图像预处理, F_{bag} 表示进行视频包特征提取, K_{atten} 表示进行注意力机制, $Pool$ 表示进行包级池化, 将视频包特征映射为视频包异常得分.

2.1　视频包特征提取

动作的发生发展具有先后顺序, 其中蕴含了大量的信息. 参照文献 [10] 中对多帧序列时空特征的提取方法, 首先对图像进行空间特征提取, 之后利用 ConvLSTM 针对时流中同一特征进行状态随时间迁移变化过程的学习. 为了得到更有效的视频包特征表示, 将视频包中提取到的各级时空特征进行加权融合编码, 从而将帧级特征转换为视频包级别的特征表示.

2.1.1　单帧空间信息提取

为了获取单帧像素之间的空间联系及纹理结构, 采用 “卷积−正则−激活”结构[16] 实现空间信息提取. 首先利用不同权重的卷积核对输入数据进行线性变换, 减少数据冗余的同时提取图片结构信息; 卷积核的权重由训练得出. 利用批规范化操作通过对隐藏层输出的自适应重参数化来防止梯度弥散效应, 加快模型收敛速度, 同时降低模型过拟合风险. 最后利用激活函数使网络模型具有非线性性能, 以便更好地拟合数据分布.

在具体的网络实现上, 由于视频包包含多帧连续图像, 为了不破坏连续帧之间的时间信息, 我们在时间维度上采用两组 “卷积−正则−激活”结构提取单帧空间信息. 对于视频包中的每一帧固定为 224×224 像素大小的图片而言, 首先经过含 128 个尺寸为 5×5 的卷积核、步长为 3 的卷积层, 得到了 128 个 75×75 像素大小的特征图, 在一个批次 (Batch) 中对不同图片的相应特征图的值进行归一化操作之后, 采用 Relu 激活函数使网络具备非线性性能; 之后经过含 64 个尺寸为 3×3 的卷积核、步长为 2 的卷积层, 得到 64 个大小为 38×38 像素的特征图, 同样以一个 Batch 为单位进行归一化操作并进行非线性映射. 正则化和激活结构不会改变特征图的尺寸. 输出特征图尺寸计算式为

$$L_{\text{out}} = \left\lceil \frac{L_{\text{in}} - W_{\text{kernel}}}{S_{\text{stride}}} \right\rceil + 2 \tag{2}$$

其中, L_{out}, L_{in} 是特征图某维度 (长、宽) 长度, W_{kernel} 指卷积核在对应尺寸上的大小, S_{stride} 为步长值, $\lceil\ \rceil$ 为向上取整.

2.1.2　多帧时间信息提取

对于图像序列而言, 每帧像素间具有很强的相关性, LSTM 只能保留其时间维度上的关联而不能很好地保留空间特征. 针对这一缺陷, 文献 [11] 中加入卷积操作, 使得对图像序列特征的提取更加有效. 具体为

$$\begin{aligned} C_t = {} & \sigma(W^3_{f1} * [h_{t-1}, x_t, C_{t-1}]) \otimes C_{t-1} + \\ & \sigma(W^3_{f2} * [h_{t-1}, x_t, C_{t-1}]) \otimes \\ & \tanh(W^2_{f3} * [h_{t-1}, x_t]) \end{aligned} \tag{3}$$

其中, 符号 $*$ 表示卷积操作. 记忆单元 C_t 记录了从起始时刻到时刻 t 的累积状态信息, h_t 表示输出的时刻 t 的最终状态, x_t 表示 t 时刻的输入, 符号 $\otimes$ 表示 Hadamard 乘积, σ 和 tanh 分别表示 sigmoid 和 tanh 激活函数, 用于对提取的信息进行非线性变换. 卷积类型 1 W^2 与卷积类型 2 W^3 的不同之处在于其输入数据有两种或者三种, 用上标表示输入数据种类数; 为了区别相同类型的不同卷积核, 在式 (2) 和式 (3) 中采用权重 W 的不同下标来表示不同的卷积核, 如 W^2_{f1} 和 W^2_{f2} 表示输入是两路数据但值不同的两个卷积核. ConvLSTM 利用 σ 选择性舍弃遗忘门输入信息, 输入门利用与遗忘门相似的结构选择性地保留经过 tanh 非线性变换的上一时刻输出与当前时刻输入的加权融合信息.

当前时刻记忆单元状态是由经过遗忘门选择性保留的上一时刻记忆单元状态与输入门输出信息相加, 当前时刻输出状态是遗忘门选择性保留的当前时刻记忆单元状态经过 tanh 非线性变换的信息, 具体实现为

$$h_t = \sigma(W^3_{f4} * [h_{t-1}, x_t, C_t]) \otimes \tanh(C_t) \tag{4}$$

本文采用的 ConvLSTM 作用的对象是经过卷积层提取过的空间特征图. 针对视频包序列中每帧对应的空间特征提取其随时间变化信息, 一则可以减少图像数据冗余, 二则可以针对关键特征追踪其时流信息. 本文采用了两层 ConvLSTM 级联来提取时流信息, 第 1 层采用 64 个 3×3 的卷积核, 第 2 层采用 32 个 3×3 的卷积层. 卷积操作时保持特征图尺寸不变, 因此两层 ConvLSTM 的输入特征图分别为 $38 \times 38 \times 64$ 和 $38 \times 38 \times 32$.

为了更好地提取视频包的空间和时间信息、高层与低层特征, 在 ConvLSTM 层后利用三维卷积层在时间和空间两个维度上对视频包多个特征图进行可训练参数的加权融合, 输出结果作为最终的视

频包的特征表示. 融合层作用于时间维度, 将 8×38 × 38 × 32 的特征图转换为1 × 38 × 38 × 32 的特征图. 融合层特征输出结果如图 2 所示.

图 2(a) 为正常帧视频包输入网络后融合层输出的部分特征图. 每个特征图细化了静态背景的建模, 突出显示运动物体区域. 图 2(b) 为异常帧视频包融合层输出特征图, 由于快速运动导致的时间流和空间流的信息重分配, 使得前景运动区域与背景静态区域融合度更高, 区分度降低.

2.2 视频包特征提取

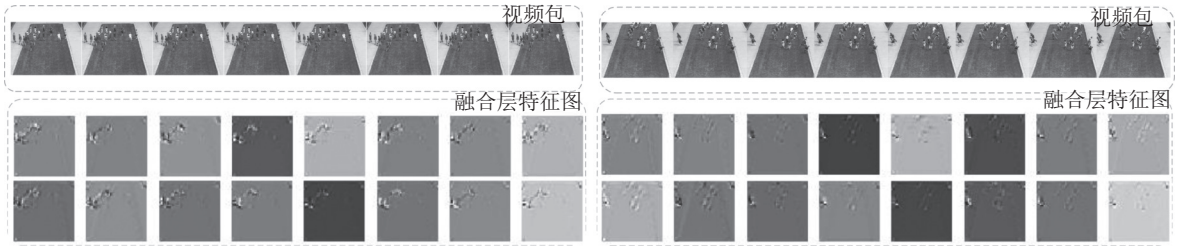
视频包内各特征元素对于最终视频包标签的影响是不同的, 因此我们考虑利用注意力机制对不同特征分配不同权重, 使关键特征对结果影响占比更大. 首先进行特征空间转换, 使原始视频包特征编码更有表现力; 之后利用可训练权值矩阵计算隐藏特征, 利用隐向量训练拟合出该特征元素的影响占比; 最后将编码特征按占比重取值, 利用包级池化操作将视频包重定义特征映射为该视频包的异常得分. 具体流程如图 3 所示.

为了得到更加紧凑的视频包特征表示, 将时空编码器输出的初级视频包特征经由两个带 Dropout 的全连接层处理, 即图 3 中的 “特征空间转换” 过程. 两层全连接层输出维度均为 512, 采用 Relu 激活函数; Dropout 参数设为 0.5, 即每次训练时有 50% 的神经单元被舍弃不参与训练, 此举可以降低神经网络过拟合的风险. 经过特征空间转换, 得到 512 维的视频包最终特征向量.

之后利用可训练权值矩阵操作将视频包最终特征转换为隐藏特征, 隐藏层特征维度设为 128. 基于隐藏特征即隐向量计算出权值向量. 为了增加变换的非线性性能, 在最终生成权值向量前引入非线性 Softmax 变换. 权值向量与视频包最终特征向量的 Hadamard 乘积即为加权视频包特征向量. 记视频包特征总数为 K, 第 k 个视频包特征为$\boldsymbol{\alpha}_k$, 可训练权值矩阵 V 将$\boldsymbol{\alpha}_k$ 转换为隐层特征; $\Psi_{\rm NL}$ 为隐层特征加入非线性操作, 本文采用 tanh 非线性函数实现; W 将非线性隐层特征转换回原始空间特征; 之后利用 softmax 操作进行非线性映射得到特征权值$\boldsymbol{\kappa}_k$, 即

$$\boldsymbol{\kappa}_k = \frac{\exp(W^{\rm T}\Psi_{\rm NL}(V\boldsymbol{\alpha}_k^{\rm T}))}{\sum\limits_{i=1}^{K}\exp(W^{\rm T}\Psi_{\rm NL}(V\boldsymbol{\alpha}_i^{\rm T}))} \tag{5}$$

经过注意力处理过后的视频包特征向量$\tilde{\boldsymbol{\alpha}}$ 见



(a) 正常视频包
(a) Normal video bag

(b) 异常视频包
(b) Abnormal video bag

图 2 融合层特征输出结果图

Fig. 2 The feature map of fusion-layer

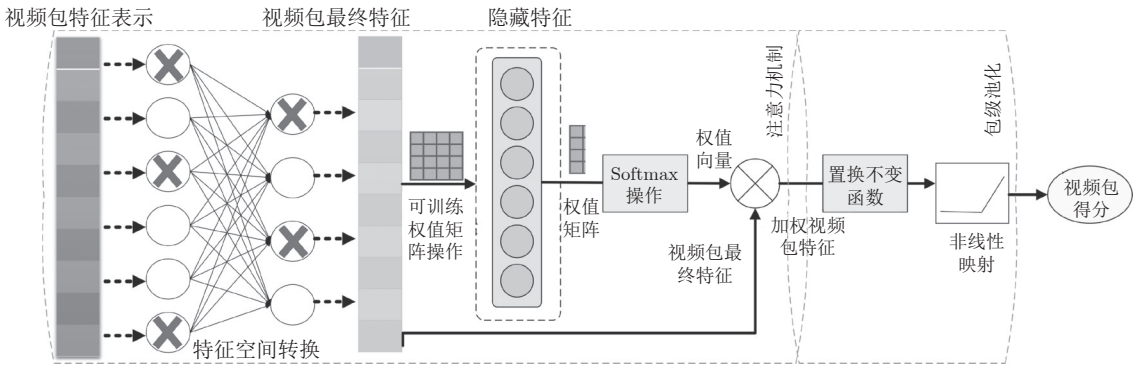


图 3 视频包得分计算流程

Fig. 3 The flowchart of bag-score calculation

式 (6). 注意力操作仅改变了特征矢量数值, 并未改变矢量维度, 因此运用了注意力机制后的矢量维度依然为 512.

$$\tilde{\boldsymbol{\alpha}} = \boldsymbol{\kappa} \otimes \boldsymbol{\alpha} \tag{6}$$

最后对加权视频包特征进行包级池化. 借助神经网络训练置换不变函数求得视频包的异常得分. 将正样本即包含异常的视频包标记为 1, 仅包含正常视频帧的视频包标记为 0, 因此可将得出的概率作为视频包得分的归一化最终结果. 至此最终实现视频包特征到视频包得分的映射. 记 Π 是对第 k 个加权视频包特征向量 $\tilde{\boldsymbol{\alpha}}_k$ 进行的置换不变操作, 在实验中取 $\tilde{\boldsymbol{\alpha}}$ 各元素累加和. σ 是 sigmoid 函数, 即图 3 中的非线性映射. $\tilde{\boldsymbol{\alpha}}$ 对应的视频包得分 B 为

$$B = \sigma(\Pi\tilde{\boldsymbol{\alpha}}_k) \tag{7}$$

2.3　损失函数

网络最终的输出是使用 sigmoid 处理过的 [0, 1] 数值, 该输出既可以作为输入被判为正类的概率, 也可作为视频包的最终异常得分. 因此损失函数的构成也从这两方面进行考虑. 根据 sigmoid 函数的计算式, 得

$$f(\Pi\tilde{\boldsymbol{\alpha}}_k) = \frac{1}{1+\mathrm{e}^{-\Pi\tilde{\boldsymbol{\alpha}}_k}} \tag{8}$$

对于输入 $\Pi\tilde{\boldsymbol{\alpha}}_k$ 求输出 B 的极大似然估计 $\hat{B}$, 其概率函数必然包含指数部分. 为了简化计算, 从信息量的角度考虑即求熵操作, 得到的二分类交叉熵损失函数 $loss_{bc}$ 为

$$loss_{bc} = \frac{1}{N}\sum_{i=1}^{N}(-B_i \lg \hat{B}_i - (1-B_i)\lg(1-\hat{B}_i)) \tag{9}$$

其中, N 为样本总数, B_i 为第 i 个样本真实值, $\hat{B}_i$ 为第 i 个样本预测值.

交叉熵损失函数在输入噪声时也会得到非 0 即 1 的结果, 会增加网络过拟合的风险, 在实际应用中不好进行阈值的选取. 考虑到视频包的特性, 我们希望包含异常帧的视频包的异常得分要比仅含正常帧的异常得分高, 因此损失函数最好可以增大类别间距. 本文加入二分类铰链损失函数 $loss_{bh}$, 即

$$loss_{bh} = \frac{1}{N}\sum_{i=1}^{N}\max(0,\, 1-2\hat{B}_i \times B_i + \hat{B}_i) \tag{10}$$

$loss_{bh}$ 通过增大正样本预测概率, 减小负样本预测概率来增大类间距, 使网络在应对少量噪声输入时有一定的容错量. 最终的网络训练损失函数为

$$loss = loss_{bc} + \lambda loss_{bh} \tag{11}$$

其中, λ 为 $loss_{bh}$ 的权重参数, 作为基础分类网络损失函数 $loss_{bc}$ 的补充项协助训练模型参数. 为了探究 λ 对模型训练的影响, 取其值为 0.5, 1, 2, 3 时分析损失函数整体性能, 结果如图 4 所示.

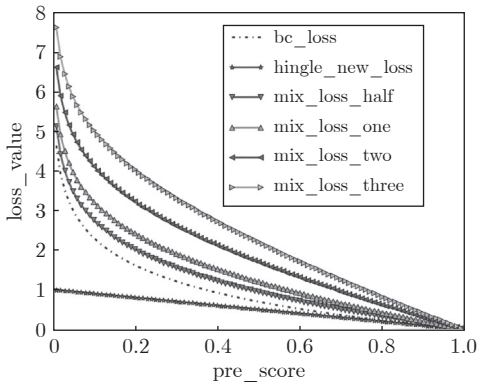


图 4　不同预测分值下的 loss 变化

Fig. 4　The loss under different predictions

取正样本的不同预测值得到不同的损失值, 网络输出 sigmoid 函数归一化 [0, 1] 值. 交叉熵损失 (bc_loss) 是分类网络的基础损失函数, 其在整个预测值空间呈非线性变化, 预测值与真实值差距越大, 惩罚越严重. 变种铰链损失 (hingle_new_loss) 意图使正样本得分比负样本得分高, 使模型更多关注整体的分类效果. 对变种铰链损失取不同权值组合成混合损失 (mix_loss), 从结果图可以看到, 当 $\lambda = 0.5$ 时, 在基础损失函数上添加的线性因素效果不甚明显; 而当 λ 取值较大时, 过多的线性因素会遮盖基础损失函数的非线性性能. 从损失函数值域考虑, 两项值域相近, 权值加倍反而会导致值域的不平衡. 综合考虑, 后续实验将采用 $\lambda = 1$ 时训练的模型, 结合非线性损失函数和线性损失函数, 增大惩罚值, 使网络收敛更快.

3　实验

3.1　数据集

本文算法应用场景为行人正常运动场景, 所要进行的异常事件检测包括跑动、滑板、车类及高空抛物. 异常事件的定义基于以下考虑: 1) 行人步行群体中驾驶交通工具属于不规范行为, 容易产生交通事故; 2) 正常人群中的跑动意味着某种紧急事情的发生, 需要引起重视; 3) 高空抛物有误伤人的嫌疑, 判为异常. 所使用的视频由固定角度的监控摄像头拍摄. 目前常用的数据集有 Avenue, UMN, UCSD.

Avenue 数据[17] 包含 CUHK 校园大道拍摄的

正常行人行走的视频, 其中训练集包含 16 个视频段共 15 328 帧, 测试集包含 21 个视频共 15 324 帧. 异常场景是人物跑动、反方向走动、闲逛等. UMN 数据集[18] 包含 11 个视频段, 异常事件类型为人群突散, 长度在 70 ~ 200 帧不等. 室内场景光照不足而视频整体偏暗. UCSD 数据集[19] 的异常事件包括骑自行车、滑冰、小型手推车及横穿人行道践踏草地. Ped1 是朝向和远离摄像头方向的片段, 有 34 个训练视频和 36 个测试视频. Ped2 是行人运动轨迹与摄像头平行的场景, 包括 16 个训练视频和 12 个测试视频.

以上数据集中训练集只包含正常行人行走的视频, 本文网络训练同时依赖正样本与负样本. 鉴于此, 本文将以上数据集进行合并筛选, 将含有异常事件的视频重新进行整合分配, 划分训练集与测试集; 同时根据本文所定义的任务, 将原有数据集中规定的不符合本文任务目标的异常事件去除, 即重新标定 ground truth. 对于整合的数据集而言, 选取事件较为单一的 40 个视频文件作为训练集, 其余 33 个视频文件为测试集. 新的数据集相较于原有数据集而言, 增加了不少挑战: 1) 场景多. 共有室内和室外六种不同场景, 不同场景摄像头位置与角度不同, 帧率及清晰度也有很大差异; 2) 异常事件类型多. 涉及单人和群体异常事件, 对于滑板、自行车、跑动等异常事件类型, 人物外观与正常行走状态相差无几, 增加了算法检测难度. 后续实验在该数据集上进行.

3.2 数据处理

本文所提出的神经网络在网络输入格式方面尚未做多尺度输入设计, 所以在原始视频输入网络时需要进行预处理. 对于输入的三通道彩色图将其转换为 224×224 像素大小的单通道灰度图. 为了降低因噪声而产生的图像像素离群值的干扰, 同时为得到图像稀疏性表示, 对输入图像矩阵进行归一化处理, 并将值域限制在 [0, 1].

3.3 实验结果

训练及测试代码运行在 Centos7 系统下, 使用 Intel i5-8600K@3.60 GHz × 6 及 NVIDIA GeForce TITAN X. 算法搭建使用以 tensorflow 为后端的 keras 框架, 语言为 python, 此外还使用了 opencv 等扩展包.

3.3.1 训练参数变化

本文算法训练使用 Adam 优化器, 学习率为 0.0001; batch_size 为 32; 当 loss 连续 5 个训练批次 epoch 不变结束训练. 网络输入视频包, 由于采用 LSTM 提取时间特征, τ 值不宜过小, 因此参考文献 [10] 中取 $\tau = 8$, 另取 $\tau = 10$ 分别进行网络的训练, 训练损失变化如图 5.

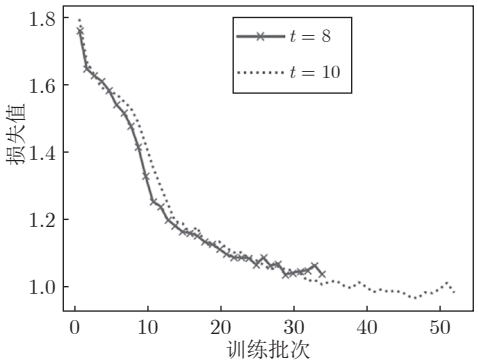


图 5 损失训练变化曲线图

Fig. 5 The loss curve in training stage

当 $\tau = 8$ (train_loss_t8) 时, 每个 epoch 含有 15 618 组输入, 在第 34 个 epoch 时训练终止, 以该处模型为此次训练代表模型, 在测试集进行测试, 所得到 AUC_T8 = 0.754, EER_T8 = 0.292. 当 $\tau = 10$ (train_loss_10) 时, 每个 epoch 含有 15 538 组输入, 训练到第 53 个 epoch 终止, 由于训练结尾模型训练损失产生波动, 取相对平稳状态下的模型, 即第 48 个 epoch 时训练的模型作为此次训练的代表模型, 在测试集测试结果为 AUC_T10 = 0.724, EER_T10 = 0.347.

从训练过程来看, 训练前期 $\tau = 8$ 模型网络损失值下降速度相对较快, 模型收敛速度更快, 而在训练终点处所达到的网络损失值相对略高; 从测试集测试定量指标来看, $\tau = 8$ 模型有较高的检测正确率, 可以预想到, 虽然 $\tau = 10$ 模型中视频包中包含了更多时间信息, 但是由于数据维度增加, 模型拟合困难, 从而出现指标降低的现象. 后续实验中本文采用 $\tau = 8$ 的第 34 个 epoch 训练的模型进行客观指标的对比测试.

3.3.2 测试结果对比与分析

1) 客观指标 AUC 及 EER

参考文献 [19–20], 本文同样使用操作者操作特征 (Receiver operating characteristic, ROC) 曲线及其所对应的面积 (Area under the curve, AUC) 来作为模型特性的衡量指标. ROC 横轴负正类率 (False positive rate, FPR) 表示负类被预测为正类的概率, 纵轴真正类率 (True positive rate, TPR) 表示正类预测正确的概率. 理想目标是 TPR = 1, FPR = 0, 故 ROC 曲线越靠拢点 (0, 1) 越好. 定义

错误率 EER = ((1-TPR)+FPR)/2.

本文算法 (Milfusion) 与时空自编码器 (Encoder)[10]、变分编码器 (Vae)[12]、多实例排名框架 (Mir)[14] 在 ROC 指标的对比结果如图 6 所示. 从图中可以看出, 本文算法曲线最接近点 (0, 1), 曲线下面积最大, 时空自编码算法次之, 变分自编码算法表现最差.

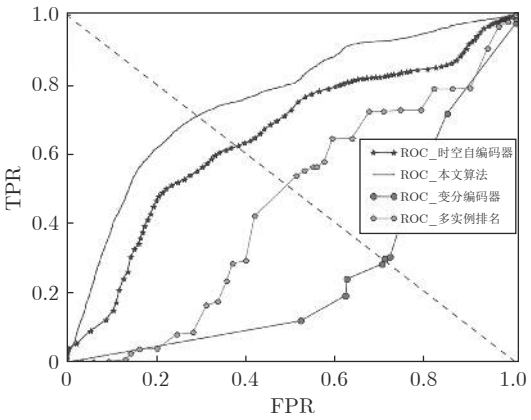


图 6　异常检测算法 ROC 曲线图

Fig. 6　The ROC curve of different algorithms

AUC 及 EER 指标见表 1, 本文算法在这两项指标上效果最好, 错误率最低.

2) 帧级正确率与事件正确率

AUC 计算时须使待检测样本中同时包含正例和负例. 而对于一些只包含异常或正常的视频而言, 该指标无法使用, 因此增加新的指标评估模型性能. 记正例判断错误的比率为帧级漏检率, 负例判断正确的比率为帧级虚警率, 二者之和为帧级错误率 (Frame-error rate, FER).

表 1　异常检测算法 AUC 及 EER 指标

Table 1　The AUC and EER of different algorithms

算法	AUC	EER
时空自编码器[10]	0.644	0.380
变分编码器[12]	0.269	0.706
多实例排名[14]	0.445	0.488
本文算法	0.754	0.292

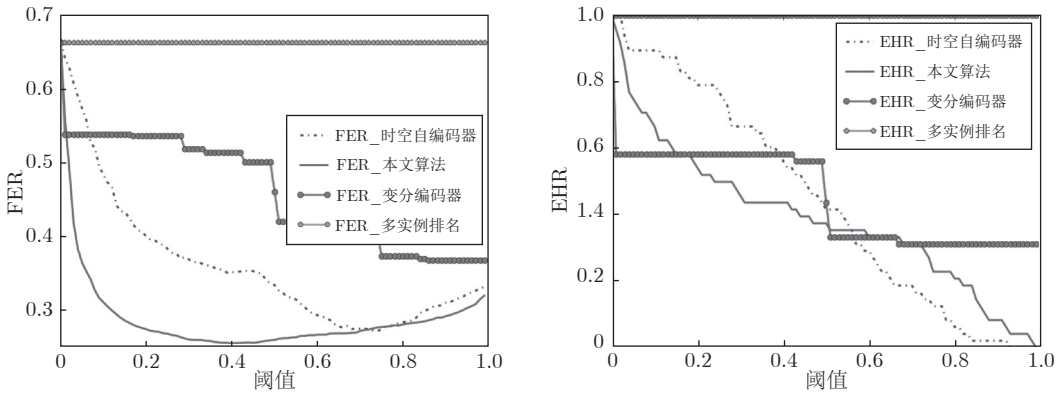
由于不同帧率人物运动速度不尽相同, 对事件发生的边缘定界有很大的歧义. 因此补充事件级别的指标, 事件识别率 (EHR), 根据检测出异常帧的百分比来判断是否检测出该异常. 事件级别的指标优势在于事件边界界定影响小. 对于定义的异常事件, 相差 20 帧以上记为两事件.

图 7 选取了 [0, 1] 范围内以 0.01 的步长作为阈值, 在该阈值下几种算法在整个数据集的 FER、EHR. 对于一个事件, 以帧级检测率不小于 60% 判为检测出该事件. 从图中可以看出, 多实例排名框架的事件检测率和帧级错误率最高, 综合来看, 其性能不佳. 本文算法的帧级错误率明显低于其余三种算法, 事件检测率指标略低于时空自编码算法. 因此, 在整体指标上本文算法具有很大优势.

3) 视频检测效果及时间对比

从上述指标来看, 本文算法和时空自编码算法性能最好, 在此仅列举本文算法和时空自编码算法的效果. 图 8 选取了 3 个视频的检测结果.

图 8(a) 来自于 UMN 数据集, 异常事件为人群突散: 本文算法在正常片段有三次持续时间较短的尖峰, 在异常片段分数较高且持续时间长; 而时空自编码算法在正常片段得分高. 图 8(b) 来自于 Avenue 数据集, 异常事件为高空抛物: 本文算法与对



(a) 帧级错误率
(a) Frame error rate

(b) 时间识别率
(b) Event hit rate

图 7　异常检测算法在帧级及事件级指标对比图

Fig. 7　The frame-level and event-level index of different algorithms

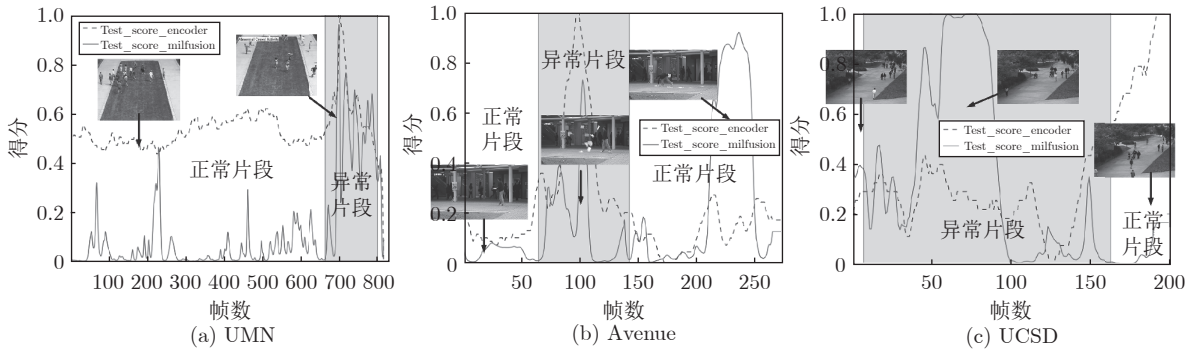


图 8 视频检测结果

Fig. 8 The results of abnormal behavior detection in videos

比算法都能正确检测出异常事件, 本文算法异常得分略低; 在人物俯身捡拾坠落物时, 本文算法出现了虚警. 图 8(c) 来自 UCSD 数据集, 异常现象为骑车: 对比算法出现漏报和虚警, 但本文算法能够检测出该事件.

在检测时间上, 本文将视频输入到输出结果耗时也进行了对比, 该时间也包含了图像尺寸改变所消耗的时间, 具体见表 2. 表中数据是在测试集上所有视频从网络输入到得到输出结果的每帧耗时的平均值. 本文网络处理时间要比时空自编码算法要短.

表 2 算法处理时间 (CPU) (ms)

Table 2 The processing time of algorithms (CPU) (ms)

时空自编码器	变分编码器	本文算法
238	245	173

本文在输出形式有很大的优势. 自编码算法输出以重建损失为基础进行分数的正则化处理, 需要将所有视频帧处理完毕后才能得到视频每帧的异常得分. 而本文算法网络输出即为视频帧的异常得分, 不需要经过特殊处理, 在使用场景方面有很大优势.

4 结束语

本文提出了一个端到端的异常检测网络, 用于监控视频行人步行群体中剧烈运动的检测. 该网络以视频包为输入, 有利于保存视频本身的时序信息及图像的纹理结构信息. 通过时空编码器充分提取视频时空特征后, 再利用注意力机制对提取到的特征进行加权处理, 突出重要特征信息, 弱化无关信息对检测结果的干扰. 最后采用包级池化将视频包级别的特征映射为视频包对应的异常得分. 该网络在输入形式上以滑动步长为 1 的窗口进行视频包的归类提取, 可以应对视频流等实时输入情况, 应用场景得到了极大地扩展; 在输出形式上, 该网络直接输出视频帧的异常得分, 不需要再做其他处理, 方便使用. 但是本文网络的缺点在于在最后得到的是异常得分而非正常或异常的分类结果, 在将得分进行类别映射时需要设置阈值, 阈值选取会极大地影响类别映射结果.

References

1 Xiao T, Zhang C, Zha H B, Wei F Y. Anomaly detection via local coordinate factorization and spatio-temporal pyramid. In: Proceedings of the 12th Asian Conference on Computer Vision. Singapore, Singapore: Springer, 2015. 66−82

2 Reddy V, Sanderson C, Lovell B C. Improved anomaly detection in crowded scenes via cell-based analysis of foreground speed, size and texture. In: Proceedings of the 2011 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops. Colorado Springs, CO, USA: IEEE, 2011. 55−61

3 Xiao Jin-Sheng, Zhu Li, Zhao Bo-Qiang, Lei Jun-Feng, Wang Li. Block-based video noise estimation algorithm via principal component analysis. *Acta Automatica Sinica*, 2018, **44**(9): 1618−1625
(肖进胜, 朱力, 赵博强, 雷俊锋, 王莉. 基于主成分分析的分块视频噪声估计. 自动化学报, 2018, **44**(9): 1618−1625)

4 Luo Hui-Lan, Wang Chan-Juan. An improved VLAD coding method based on fusion feature in action recognition. *Acta Electronica Sinica*, 2019, **47**(1): 49−58
(罗会兰, 王婵娟. 行为识别中一种基于融合特征的改进VLAD编码方法. 电子学报, 2019, **47**(1): 49−58)

5 Xiao J, Shen M, Lei J, Zhou J, Klette R, Sui H. Single image dehazing based on learning of haze layers. *Neurocomputing*, 2020

6 Zhou Y Z, Sun X Y, Zha Z J, Zeng W J. MiCT: Mixed 3D/2D convolutional tube for human action recognition. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, Utah, USA: IEEE, 2018. 449−458

7 Ionescu R T, Khan F S, Georgescu M I, Shao L. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video. In: Proceedings of the 2019 IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA: IEEE, 2019. 7834−7843

8 Cai Rui-Chu, Xie Wei-Hao, Hao Zhi-Feng, Wang Li-Juan, Wen Wen. Abnormal crowd detection based on multi-scale recurrent neural network. *Journal of Software*, 2015, **26**(11): 2884−2896
(蔡瑞初, 谢伟浩, 郝志峰, 王丽娟, 温雯. 基于多尺度时间递归神经网络的人群异常检测. 软件学报, 2015, **26**(11): 2884−2896)

9 Yuan Fei-Niu, Zhang Lin, Shi Jin-Ting, Xia Xue, Li Gang. Theories and applications of auto-encoder neural networks: a literature survey. *Chinese Journal of Computers*, 2019, **42**(1): 203−230

(袁非牛, 章琳, 史劲亭, 夏雪, 李钢. 自编码神经网络理论及应用综述. 计算机学报, 2019, **42**(1): 203−230)

10 Chong Y S, Tay Y H. Abnormal event detection in videos using spatiotemporal autoencoder. *International Symposium on Neural Networks*, 2017, (10262): 189−196

11 Shi X J, Chen Z R, Wang H, Yeang D Y. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In: Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Quebec, Canada: MIT Press, 2015. 802−810

12 An J, Cho S. Variational autoencoder based anomaly detection using reconstruction probability. *SNU Data Mining Center*, Korea, Spe. Lec. on IE, 2015, **2**: 1−18

13 Yuan Jing, Zhang Yu-Jin. Application of sparse denoising auto encoder network with gradient difference information for abnormal action detection. *Acta Automatica Sinica*, 2017, **43**(4): 604−610
(袁静, 章毓晋. 融合梯度差信息的稀疏去噪自编码网络在异常行为检测中的应用. 自动化学报, 2017, **43**(4): 604−610)

14 Sultani W, Chen C, Shah M. Real-world anomaly detection in surveillance videos. In: Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, Utah, USA: IEEE, 2018. 6479−6488

15 Tran D, Bourdev L, Fergus R, Torresani L, Paluri M. Learning spatiotemporal features with 3D convolutional networks. In: Proceedings of the 2015 IEEE International Conference on Computer Vision. Santiago, Chile: IEEE, 2015. 4489−4497

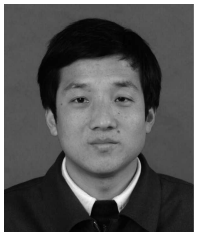
16 Xiao Jin-Sheng, Zhou Jing-Long, Lei Jun-Feng, Liu En-Yu, Shu Cheng. Single image dehazing algorithm based on the learning of hazy layers. *Acta Electronica Sinica*, 2019, **47**(10): 2142−2148
(肖进胜, 周景龙, 雷俊锋, 刘恩雨, 舒成. 基于霾层学习的单幅图像去雾算法. 电子学报, 2019, **47**(10): 2142−2148)

17 Lu C W, Shi J P, Jia J Y. Abnormal event detection at 150 FPS in MATLAB. In: Proceedings of the 2013 IEEE International Conference on Computer Vision. Sydney, NSW, Australia: IEEE, 2013. 2720−2727

18 Unusual crowd activity dataset of University of Minnesota [Online], available: http://mha.cs.umn.edu/Movies/Crowdctivity-All. avi, October 25, 2006

19 Mahadevan V, Li W X, Bhalodia V, Vasconcelos N. Anomaly detection in crowded scenes. In: Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Francisco, CA, USA: IEEE, 2010. 1975−1981

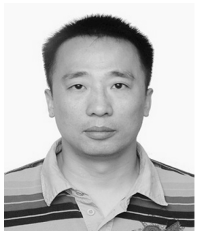
20 Saligrama V, Chen Z. Video anomaly detection based on local statistical aggregates. In: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition. Providence, RI, USA: IEEE, 2012. 2112−2119



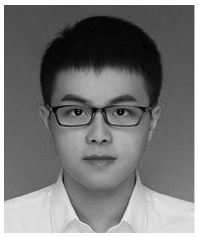
肖进胜 博士, 武汉大学电子信息学院副教授. 2001 年于武汉大学获理学博士学位. 主要研究方向为视频图像处理, 计算机视觉.
E-mail: xiaojs@whu.edu.cn
(**XIAO Jin-Sheng** Ph.D., associate professor at the School of Electronic Information, Wuhan University. He received his Ph.D. degree from Wuhan University in 2001. His research interest covers video and image processing, computer vision.)

申梦瑶 武汉大学电子信息学院硕士研究生. 2018 年获得武汉大学电子信息学院工学学士学位. 主要研究方向为视频图像处理, 计算机视觉.
E-mail: shenmy@whu.edu.cn
(**SHEN Meng-Yao** Master student at the School of Electronic Information, Wuhan University. She received her bachelor degree from Wuhan University in 2018. Her research interest covers video and image processing, computer vision.)

江明俊 武汉大学电子信息学院硕士研究生. 2019 年获得武汉大学电子信息学院工学学士学位. 主要研究方向为视频图像处理, 计算机视觉.
E-mail: 2015301200236@whu.edu.cn
(**JIANG Ming-Jun** Master student at the School of Electronic Information, Wuhan University. He received his bachelor degree from Wuhan University in 2019. His research interest covers video and image processing, computer vision.)



雷俊峰 博士, 武汉大学电子信息学院副教授. 2002 年于武汉大学获得理学博士学位. 主要研究方向为视频图像处理, 计算机视觉. 本文通信作者.
E-mail: jflei@whu.edu.cn
(**LEI Jun-Feng** Ph.D., associate professor at the School of Electronic Information, Wuhan University. His research interest covers video and image processing, and computer vision. Corresponding author of this paper.)



包振宇 武汉大学电子信息学院硕士研究生. 2018 获得武汉理工大学信息工程学院工学学士学位. 主要研究方向为视频图像处理, 计算机视觉.
E-mail: 2018282120154@whu.edu.cn
(**BAO Zhen-Yu** Master student at the School of Electronic Information, Wuhan University. He received his bachelor degree from Wuhan University of Technology in 2018. His research interest covers video and image processing, and computer vision.)