计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200900093

在线异常事件检测的时序建模

卿来云[1]　张建功[1]　苗　军[2]
1 中国科学院大学计算机科学与技术学院　北京 100049
2 北京信息科技大学网络文化与数字传播北京市重点实验室　北京 100101

摘　要　弱监督异常事件检测是一项极富挑战性的任务,其目标是在已知正常和异常视频标签的监督下,定位出异常发生的具体时序区间。文中采用多示例排序网络来实现弱监督异常事件检测任务,该框架在视频被切分为固定数量的片段后,将一个视频抽象为一个包,每个片段相当于包中的示例,多示例学习在已知包类别的前提下训练示例分类器。由于视频有丰富的时序信息,因此重点关注监控视频在线检测的时序关系。从全局和局部角度出发,采用自注意力模块学习出每个示例的权重,通过自注意力值与示例异常得分的线性加权,来获得视频整体的异常分数,并采用均方误差损失训练自注意力模块。另外,引入LSTM和时序卷积两种方式对时序建模,其中时序卷积又分为单一类别的时序空洞卷积和融合了不同空洞率的多尺度的金字塔时序空洞卷积。实验结果显示,多尺度的时序卷积优于单一类别的时序卷积,时序卷积联合包内包外互补损失的方法在当前UCF-Crime数据集上比不包含时序模块的基线方法的AUC指标高出了3.2%。

关键词:异常事件检测;弱监督学习;多示例学习;注意力机制;时序卷积网络
中图法分类号　TP391

Temporal Modeling for Online Anomaly Detection

QING Lai-yun[1],ZHANG Jian-gong[1] and MIAO Jun[2]
1 School of Computer Science and Technology,University of Chinese Academy of Sciences,Beijing 100049,China
2 Beijing Key Laboratory Internet Culture Digital Dissemination Research,Beijing Information Science & Technology University,Beijing 100101,China

Abstract　Weakly supervised anomaly detection (WSAD) is a challenging task in that there is only normal and anomaly video label supervision but it is required to localize intervals where anomalies take place. We employ multiple instance learning (MIL) network for weakly supervised anomaly detection,which regards the input video as a bag and the segments chunked from the video as instances in it. We train the instance classifier with only label of video level (bag level),while the label of instance level is unknown. As there is strong temporal information in videos,we focus on temporal relationship for online anomaly detection in surveillance videos. We consider both global and local perspective and use self-attention module to learn each instance weight. We get the linear weighted sum of self-attention score and instance anomaly score,which represents video level anomaly score. Then the mean square error loss is employed to train the self-attention module. Online constraints allow us to use historical and current video clips only,without future frames. In order to model the temporal structure of video,we introduce LSTM and temporal convolutional network (TCN) into WSAD problem. We explore single rate dilated temporal convolutional network,and pyramid dilated temporal convolutional network (PDTCN) which fuses multi-scale feature with different rates. Experiments show that the AUC of PDTCN with complementary inner and outer bag loss is higher than that of the baseline method without temporal modeling by 3.2% on UCF-Crime dataset.

Keywords　Anomaly detection,Weakly-supervised learning,Multiple instance learning,Attention module,Temporal convolutional network

到稿日期:2020-09-13　返修日期:2020-10-25
基金项目:国家自然科学基金面上项目(61872333Y);北京未来芯片技术高精尖创新中心科研基金(KYJJ2018004);北京教委科技计划项目(KM201911232003);北京市自然科学基金(4202025)
This work was supported by the NSFC (61872333Y),Research Fund from Beijing Innovation Center for Future Chips (KYJJ2018004),Beijing Municipal Education Commission Project (KM201911232003) and Beijing Natural Science Foundation (4202025).
通信作者:卿来云(lyqing@ucas.ac.cn)

1 引言

安防作为近年来计算机视觉领域热门的研究方向，与视频分析研究有着紧密的联系。在真实的监控视频中，一个常见的需求是自动识别视频流中的异常事件，即异常事件检测任务。异常事件检测的应用范围非常广泛，可以应用于生活场景、医疗行业、金融服务、农业、交通以及教育行业等众多领域，在异常事件的突发阶段对其进行主动识别、主动跟踪和主动预警。

为了节省标注的时间和人力，弱监督学习是近年来的研究热点。监控场景中的弱监督异常事件检测（weakly supervised anomaly detection）指训练时仅需标注视频中异常或正常事件的类别信息，而无须在视频中标注异常事件的起止时间；测试时定位出视频流中异常事件发生的具体时域区间。该方法虽然在精度上会有所损失，但节约了大量的标注成本。

另外，视频监控场景下的异常事件检测一般采取在线处理的方式。与离线处理完整视频数据的传统检测方法不同，在线处理是视频流在异常事件实时发生的过程中对其进行检测并完成预警。由于在线的实时性约束，在线检测方法只能利用历史和当前的视频信息。输入一个视频流，$V=(I_1,I_2,I_3,\cdots,I_t)$，包含视频从第 1 到当前第 t 个视频片段，在线异常事件检测的目标是利用过去和当前的视频信息来判断当前视频片段是否异常。在线视频异常事件检测可以定义为最大化后验概率问题，即：

$$\underset{Y_t}{\operatorname{argmax}}\, P((Y_t \mid I_1,I_2,I_3,\cdots,I_t))$$

其中，Y_t 表示 t 时间段视频的异常类别。在线视频流的异常事件检测能连续实时检测在线视频，当一个新产生的视频帧到达时，算法无须等待后续的未来帧，可以立即对当前视频片段直接进行处理。

由于视频有很强的时序关系，有些事件（如纵火、袭击等）都是多阶段的，一些异常事件持续时间较长，因此仅仅利用单个视频片段构成的示例训练示例分类器是不充分的。考虑到异常视频检测任务的数据集以监控视频为主，其实时性要求较高，因此我们主要关注两个问题：1）融合视频的时域上下文信息；2）监控视频的在线检测。

本文融合整体和局部的信息，引入时域注意力模块，它由 3 层级联的全连接层构成。通过自注意力值与示例异常得分的线性加权，我们获得了视频的整体得分。

常见的时序建模方式还有时序卷积网络（Temporal Convolutional Network，TCN）和递归神经网络（Recurrent Neural Networks，RNN）。在 RNN 中，通常的做法是采用空域卷积神经网络（Convolutional Neural Networks，CNN）抽取每个视频帧的特征，再结合 RNN 学习视频时序历史信息，将 RNN 每一个时间点的输出特征作为分类的依据。本文中，我们对视频包内的示例序列采用 C3D 特征提取，再将其送入双向长短时记忆网络（Long Short-Term Memory，LSTM），对示例级别的时序信息进一步编码。双向 LSTM 两个分支的中间特征经过拼接后送入示例分类器，最后通过包内包外联合损失来训练该分类器。

本文受 Bai 等[1] 提出的时序模型的启发引入了时序卷积。该卷积对当前示例进行卷积操作时，卷积核只会覆盖历史和当前示例特征，无须未来示例的特征参与。这一特点非常符合在线异常视频检测的任务要求。我们对卷积核的空洞大小做了相应的实验探索，进一步应用金字塔时序卷积把多尺度的视频信息进行拼接。实验结果表明，该方法的效果更为突出。

本文的主要贡献如下：1）针对异常事件时序跨度大且可能存在多段异常的特点，从视频整体和局部角度出发引入时序注意力机制；2）考虑到监控视频流的在线特点，将 LSTM、时序空洞卷积及金字塔时序空洞卷积引入异常事件检测问题中；3）以包内包外联合损失为基础方法，结合时序卷积在 UCF-Crime[2] 数据集上获得了非常有竞争力的结果。

2 相关工作

2.1 多示例学习在弱监督任务中的应用

在图像弱监督物体检测中，图片中包含目标物体的区域构成一个正包示例，不包含目标物体的区域构成一个负包示例。在多示例学习中，只有包级的监督信息是已知的，准确的示例级的监督信息未知。多示例学习的目标是在只有包级标签的前提下训练一个示例分类器。Bilen 等[3] 提取目标提名作为示例，然后应用两个全连接层分别构建分类和检测模块。Tang 等[4] 提出了一个在线的示例分类器回归的方法，并将其集成到多示例学习网络中，用于有效地回归弱监督目标检测的目标框。一些方法也使用高质量的伪标签逐步回归目标示例的边界[5-6]。

弱监督时域动作检测指训练集中只标记整段视频包含的动作类别，并没有片段级别的时间信息。弱监督主要利用分类来进行检测，分类过程中网络会主动关注到对分类最有区分度的视频片段，从而进行动作定位。Nguyen 等[7] 将 Zhou 等[8] 的类激活图迁移到视频的时域上，并在视频片段中引入自注意力以表示某个片段为前景动作的重要性。Paul 等[9] 在分类网络的基础上提出了协同动作相似性损失，用排序损失实现在特征空间上驱使同类动作的视频特征尽可能近，不同类的视频特征相对远。Lee 等[10] 提出背景抑制网络 BaS-Net 并引入背景类，为了构造背景类的负样本，在并行的第二支网络中引入注意力模块以有效抑制背景的影响。

2.2 异常事件检测

由于异常事件发生的频率很低，因此数据的收集和标注非常困难。训练中的正样本（异常）一般远少于负样本（正常）。监控场景中正常和异常事件类内复杂多样，难以一一列举。

一种方式是所有训练集都是正常类样本，测试集包含正常类样本和异常类样本，并且需要定位异常区间。这种情况下，大致有 3 种模型可以实现异常事件监测：重构模型、预测模型和生成模型。这些模型旨在学习正常类样本的分布，异常样本离该分布距离较远。例如，在重构模型中，异常样本的重构误差较大，因此能通过重构的帧与真实帧的差异来判断是否发生了异常。Hasan 等[11] 使用卷积自编码器重构了输入帧序列，然后使用重构误差去定位异常。Lu 等[12] 使用基于字典的方法学习正常模式的信息。Zhao 等[13] 使用 3D 卷

积网络首先对视频片段进行编码以获取低维表示，再使用两个 3D 反卷积网络进行解码以构成重构和预测分支，从而得到近似的重构片段及预测片段。测试时输出的重构误差作为最终结构定位异常。Liu 等[14]使用 U-Net 作为生成对抗网络(Generative Adversarial Networks，GAN)。训练的生成器利用 GAN 使重构的预测图像逼近真实图像，利用 Flownet[15]产生光流，捕捉对应的运动信息，使重构光流逼近真实光流，生成的预测误差用于定位异常。Ionescu 等[16]通过聚类方式检测异常，该方法与无监督图片分类相结合，同时利用物体检测的方式抑制背景的影响。由于自编码器的泛化能力过强，实际上会把部分异常也进行较好的重构，因此 Gong 等[17]针对此问题对自编码器进行了记忆增强操作，尝试解决这个问题。

上述无监督方法均通过重构误差或预测误差来判断当前视频片段是否异常。但这种方法从理论上无法构造出一个所谓的正常模板，因为正常模板中包含各种各样的事件，同样也无法构造异常事件的集中表示。面对这种开放域问题，Sultani 等[2]提出了一种基于深度多示例排序网络的弱监督学习框架，同时提出了一个有监督的异常事件检测数据集，他们还认为异常事件检测应该在弱监督框架下进行学习。此处弱监督指在训练时，只知道一段视频中有或没有异常事件，而异常事件的种类以及具体的发生时间是未知的。文献[2]采取两阶段的框架，即不管异常事件的种类，首先生成异常事件的区间提名，然后对区间提名中包含的异常事件进行分类。Zhang 等[18]在多示例排序网络的基础上引入了包内损失，进一步约束了算法的搜索空间。本文亦讨论了在弱监督情况下的异常事件检测，并对其中的时序关系做了进一步探索。

3 异常事件检测的时序建模

3.1 基于多示例学习的异常事件检测

图 1 给出了本文采用的基于多示例学习的框架结构。

首先视频被无重叠地切分为固定数量的视频片段，构成包中示例，然后使用 C3D[19]抽取每个示例的特征。模型同时使用正样本包(异常)$\mathcal{B}_a$和负样本包(正常)$\mathcal{B}_n$参与训练，其中正样本包中至少有一个示例是异常的，负样本包中每个示例均为正常事件。

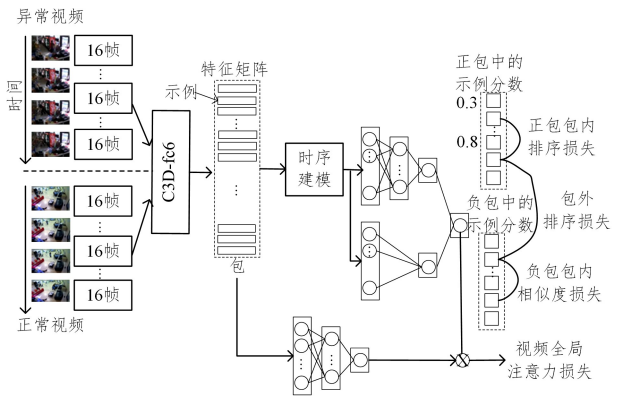

图 1 弱监督异常事件检测的整体框架图

Fig. 1 Framework of the proposed weakly supervised anomaly detection

在示例特征的基础上，本文分别从时序注意力、LSTM 以及时序卷积这 3 个方面出发，探索示例级别的时序信息。对于时序建模后的输出，本文训练了两个全连接层分支，将其所有输出的均值作为当前示例的异常得分 f。

网络的损失函数包括包内包外联合损失、注意力损失、和稀疏性约束，具体定义为：

$$\begin{aligned} l(\mathcal{B}_a,\mathcal{B}_n)=&\lambda_1\max(0,1-\max_{i\in\mathcal{B}_a}f(V_a^i)+\max_{i\in\mathcal{B}_n}f(V_n^i))+\\ &\lambda_2\max(0,1-\max_{i\in\mathcal{B}_a}f(V_a^i)+\min_{i\in\mathcal{B}_a}f(V_a^i))+\\ &\lambda_3(\max_{i\in\mathcal{B}_n}f(V_n^i)-\min_{i\in\mathcal{B}_n}f(V_n^i))+\lambda_4(y_a-\tilde{y}_a)^2+\\ &\lambda_5(y_n-\tilde{y}_n)^2+\lambda_6\textstyle\sum_i f(V_a^i)\end{aligned}$$

其中，V 表示示例，$\tilde{y}_a=1$，$\tilde{y}_n=0$ 分别表示异常视频的实际标签 1 和正常视频的实际标签 0。上述损失中，第一项是包外排序损失，表示期望的异常视频中异常事件发生的片段比正常视频的所有片段拥有更高的异常响应得分；第二项是正包包内排序损失，期望正包内的正示例(高异常得分)和负示例(低异常得分)间的距离尽可能拉开；第三项是负包包内相似性损失，由于是负包，所有片段均不含异常，因此我们迫使负包中最高和最低得分的两个示例间的距离最小；第四项和第五项是正包和负包的全局注意力损失，这里采用均方误差损失；最后一项表示稀疏损失，因为视频中的异常片段出现的个数和时间是稀疏且短暂的。

3.2 时序注意力机制

Zhu 等[20]针对异常事件检测任务首次提出了时序注意力模块，基于自注意力的模块求得视频整体的异常分数。与文献[20]中的时序注意力模块不同的是，本文将其作为异常和正常二分类的问题处理。该模块的整体结构如图 2 所示。

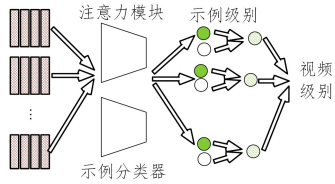

图 2 时序注意力模块

Fig. 2 Temporal self-attention module

时序注意力模块使用自注意力机制，其输入为经过特征提取的示例特征，目的是学习当前示例的重要性。由于视频流的特点，自注意力模块对每个示例均输出一个注意力值，通过加权求和计算出视频整体的异常得分。这样设计不仅考虑到了当前示例，更从全局考虑到了视频的时域信息，从而能更有效地定位视频异常片段。其计算式如下：

$$\begin{cases} f_t^1=\tanh(W_1x_t+b_1),dropout(f_t^1)\\ f_t^2=W_2f_t^1+b_2,dropout(f_t^2)\\ f_t^{att}=\tanh(W_3f_t^2+b_3)\end{cases}$$

其中，x_t 表示输入特征，f_t^1，f_t^2 分别是示例特征经过全连接后输出的中间结果，f_t^{att} 表示学到的注意力权重。最后通过将自注意力值与原特征相乘，来计算视频的全局注意力得分。

$$y_a=\sum_{i\in\mathcal{B}_a}f_t^{att}f(V_a^i)$$

$$y_n=\sum_{i\in\mathcal{B}_a}f_t^{att}f(V_n^i)$$

对于异常视频，其整体注意力得分 y_a 应该趋近异常值 1，

正常视频的整体得分y_n应该趋近正常值 0。

实验中采用的注意力模块包含级联的全连接层和非线性层以及用丢弃(dropout)减少过拟合。第一个全连接层包含 256 个节点，第二个全连接层包含 64 个节点，最后一个全连接层输出对当前示例的注意力得分。一个包含 N 个示例的视频$(x_1,x_2,x_3,\cdots,x_N)$经过特征提取和注意力模块后，可得到一个 $1\times N$ 的注意力分数向量$[f_1^{att},f_2^{att},f_3^{att},\cdots,f_N^{att}]$。根据视频的标注信息，我们可以从视频级别设计全局损失函数，实验中我们采用均方误差损失度量。对包内和包外排序损失而言，由自注意力构成的全局损失可以作为补充损失，从视频所有示例的角度出发，学习异常和正常视频之间的区别，由此训练示例分类器。

3.3 LSTM

本文进一步采用图 3 所示的双向 LSTM 对示例间的时序信息进行建模。对于输入由连续 N 个示例组成的视频序列$(x_1,x_2,x_3,\cdots,x_N)$，其中每个示例通过 C3D 网络提取特征，变成 4096 维的特征向量。正向 LSTM 层对视频序列中每个示例特征从左到右计算输出中间结果$\overrightarrow{h_t}$，反向 LSTM 层对同一个视频的每个示例从右向左计算输出中间结果$\overleftarrow{h_t}$，这种双向 LSTM，先从左到右，按时间顺序学习由过去到现在的视频序列的历史信息，然后从右向左，逆序学习视频序列从当前到过去的信息变化。通过组合$\overrightarrow{h_t}$和$\overleftarrow{h_t}$形成$[\overrightarrow{h_t},\overleftarrow{h_t}]$，这样可以有效利用视频的历史示例和当前示例，并依据历史信息对当前示例的异常情况做进一步的优化决策。

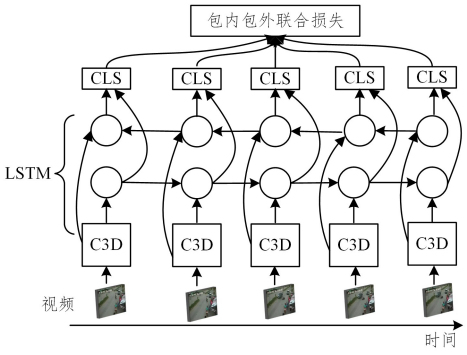

图 3 长短时记忆网络应用于弱监督异常事件检测

Fig. 3 LSTM for weakly supervised anomaly detection

3.4 金字塔时序空洞卷积

除了利用 LSTM 处理在线视频的时序问题，本文参考 Wang 等[21]关于在线动作检测的时序建模综述，将一系列时序卷积网络引入异常事件检测问题中。特别地，对于 WaveNet[22]的变种，其卷积核只会覆盖历史示例，而不需要未来示例的特征。本文评估了单一不同空洞率的时序卷积以及 Li 等[23]使用的复合金字塔时序空洞卷积在弱监督异常事件检测任务上的应用效果。

本文的目标是得到每个示例对应的异常得分，因此当输入经过时序卷积后，它的输出与输入应是等长的，这里通过填充操作来保持时序空洞卷积的等长性。对于视频示例的集合输入 $Input=(x_1,x_2,x_3,\cdots,x_N)$，时序卷积模型输出的特征如下：

$$F_o=f_{o_1}^{(r)},f_{o_2}^{(r)},f_{o_3}^{(r)},\cdots,f_{o_t}^{(r)}$$

$$f_{o_t}^{(r)}=\sum_{i=0}^{size-1}W_{[i]}^{(r)}\times x_{[t-r\times i]}$$

其中，r 是空洞率，表示做时序空洞卷积操作时对视频包内示例的时序跨度；$W^{(r)}$ 表示卷积核参数，其中 $size$ 表示时序空洞卷积核的大小；$t-r\times i$ 表示朝过去的方向做时序卷积。当 $r=1$ 时退化为正常的卷积操作，当 $r>1$ 时则表示空洞卷积操作。为有效扩大时域感受野，我们尝试了两种方式：

(1)单一不同空洞率的时序空洞卷积，主要涉及多种空洞率$r_1,r_2,r_3,\cdots,r_n$；

(2)混合金字塔时序空洞卷积。

金字塔时序空洞卷积（Pyramid Dilated Temporal Convolution，PDTC）是单一空洞率时序空洞卷积的复合版本，它将多种空洞率$r_1,r_2,r_3,\cdots,r_n$的特征拼接在一起，其复合特征的定义如下：

$$f_t^{\text{concat}}=concat(f_t^{r_1},f_t^{r_2},f_t^{r_3},\cdots,f_t^{r_n})$$

不同的空洞率 r 可覆盖多种不同的时序区间，这种做法往往比使用单一类型的时序空洞卷积更加有效。融合多尺度信息有助于检测不同时长的异常事件。经过拼接后的卷积特征被送入非线性模块，具体表达式如下：

$$f_t^{\text{concat}}=ReLU(Wf_t^{\text{concat}}+b)$$

其中，非线性层的 W 和 b 作为全连接层的参数，用于对金字塔拼接的特征进行特征降维。

后续将输出送入示例的分类器，通过包内包外联合损失来训练示例分类器，对示例的异常性打分。本文在实验中使用了 3 种空洞卷积率 $r=1,2,3$，并对它们进行组合拼接，测试模型的有效性。图 4 给出了卷积核大小为 3 的时序卷积网络的示意图。

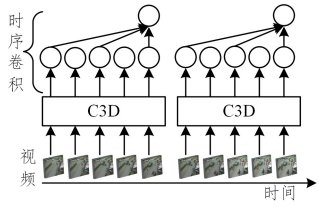

图 4 卷积核大小为 3 ($r=2$) 的时序卷积网络

Fig. 4 Temporal convolution with kernel size of 3 ($r=2$)

本文通过对上述单一类型的时序空洞卷积进行特征组合，形成复合金字塔特征。图 5 为对 3 种空洞卷积率 $r=1,2,3$ 的时序空洞卷积后的特征组合拼接示意图。本文组合了不同尺度的时序特征，降低了模型对不同时长异常事件的敏感度。相比单一类型的时序空洞卷积，复合金字塔时序空洞卷积多了一步全连接操作，将拼接后的特征降维到示例分类器的输入维度。为了验证复合金字塔特征的有效性，本文分别对空洞卷积率为 1 和 2，2 和 3，1 和 3 以及 1，2，3 的组合进行实验。

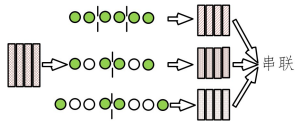

图 5 金字塔时序空洞卷积

Fig. 5 Pyramid dilated temporal convolution

4 实验结果

4.1 UCF-Crime 数据集

UCF-Crime 数据集是 Sultani 等[2]提出的监控场景下的异常事件检测数据集，可用于异常事件的识别与检测。该数据集提供了 13 种异常事件，包括虐待、逮捕、纵火、袭击、交通事故、入室盗窃、爆炸、打架、抢劫、枪击、偷窃、入店盗窃和破坏，以及日常生活中的正常事件共计 1 900 个视频，其中训练集包含 800 个正常视频以及 810 个异常视频，测试集中共有 290 个视频。其中，一些视频中包含多个异常片段。数据集的总时长为 128 h，视频平均帧数为 7 274，视频尺寸大部分是 240×320，且为 MP4 格式，帧率约为 30 FPS。绝大多数异常视频的异常片段占比为 5%～20%，且个别视频中异常事件占据总视频时长的一半以上。综合来看，异常事件持续时长的占比趋于多样化。

UCF-Crime 数据集是针对弱监督方法构建的训练集和测试集，可用于视频分类和弱监督异常事件检测任务。其中，训练集只有视频级别的标注，测试集对异常视频做了帧级标注，标注出异常发生的起始和终止帧，从而可以对测试结果进行时域异常片段的测试评估。

异常事件检测常用的评测方法是视频帧级接收者操作特征(Receiver Operating Curve，ROC)曲线，将曲线下的面积(Area Under Curve，AUC)作为评价指标。

4.2 实验设置

测试集包含 290 个视频，其中有 150 个正常视频和 140 个异常视频。实验时，本文和 Sultani 等[2]的做法类似，首先将视频固定帧率为 30 FPS，然后将每一帧放缩到 240×320 大小，经过初步处理后，我们将视频切分为 32 段不重叠的时序片段作为包中示例，为了公平比较，采用相同的特征提取方式来抽取视频的特征，其中选取 C3D 的 FC6 这一全连接层。为了获取示例特征，我们对当前视频分段内的所有 16 帧 C3D 特征取平均，将平均值作为当前视频段的特征。从而得到 32×4 096 维的特征。在训练过程中，批训练大小为 60，其中包含视频数据集中随机选取的异常视频 30 个和正常视频 30 个。接下来，经过特征提取的 4 096 维特征被并行送入时序模块和注意力模块，时序模块的输出被送入示例分类器，该分类器采用并行的双分支聚合方式求得示例最后的分类得分。我们使用 Adagrad[24]优化方法，并设置最初的学习率为 0.01，包外排序损失、异常包内排序损失和正常包内排序损失的超参系数依次设置为 $\lambda_1=\lambda_2=\lambda_3=1$，全局注意力损失的超参系数分别为 $\lambda_4=\lambda_5=0.01$。同时，根据文献[2]的做法，将稀疏项的系数设置为 $\lambda_6=8\times10^{-5}$。以上实验均在 Ubuntu16.04 系统，NVIDIA TITAN-X GPU 上运行。

4.3 实验结果及分析

网络的消融实验结果如表 1 所列。本文在 Sultani 等[2]的基线的基础上，比较了 LSTM 和单一类型时序卷积模型的结果。我们发现，LSTM 的结果(0.35%)比 $r=1$ 时序卷积(1.71%)差，原因可能是 C3D 特征已经蕴含了时序信息，其中每个示例由 16 帧组成，进一步跨示例的长时序建模效果没有大幅提升，即 C3D 和 LSTM 不太兼容，可能帧级的 CNN 和光流特征比 C3D 特征与 LSTM 的组合更具亲和性。$r=1,2$ 的单一类型的时序空洞卷积分别比 Sultani 等的基线方法高出了 1.71%，0.86%，但当 $r=3$ 时，性能比 Sultani 等的基线方法降低了 0.08%。由于 C3D 已经融合了时域信息，我们发现，对于单一类型的时序空洞卷积，随着卷积的尺寸增大，长时序 C3D 特征异常事件的检测精度成下降趋势。其中，邻接示例与当前示例一起卷积($r=1$)的效果最好。融合了多尺度信息的空洞率为 1 和 2 的混合金字塔时序空洞卷积的效果更优。

表 1 消融实验结果

Table 1 Experimental results of ablation study

方法	AUC
基线(Sultani 等[2])	75.41
基线(包内包外联合损失)	77.60
注意力	77.95
PTCN($r=1$)	78.12
PTCN($r=1$&$r=2$)	**78.96**

图 6 给出了本文方法与一些现有方法的结果。本文方法在 UCF-Crime 数据集上优于没有加入时序模块的基线方法[18] 1.3% (78.96% vs. 77.60%)。

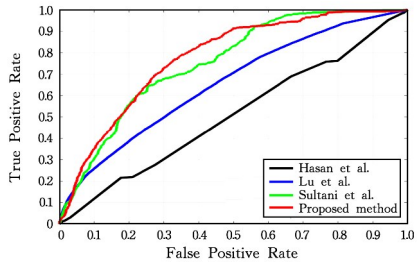

图 6 UCF-Crime 数据集上各类方法的 ROC 比较

Fig. 6 ROC on the UCF-Crime dataset

表 2 列出了本文模型与不同方法在 UCF-Crime 数据集上的 AUC 对比结果。

表 2 UCF-Crime 数据集上各类方法的 AUC 比较

Table 2 AUC on the UCF-Crime dataset

方法	AUC
基线(包内包外联合损失)[18]	77.60
Hasan 等[11]	50.60
Lu 等[12]	65.51
Sultani 等[2]	75.41
Zhu 等[20]	79.0
本文方法	**78.96**

观察图 7 中的实验结果可以发现，大多数异常类的 AUC 均显示出了不同程度的提升，其中袭击、抢劫、入店盗窃、纵火和破坏的涨幅最大，分别高出 22%，14%，11%，9%和 8%。其中逮捕、打架、爆炸、枪击和入室盗窃有小幅增长，分别增长了 6%，5%，4%，4%和 2%。然而，偷窃和交通事故有小幅下降，通过分析发现，偷窃需要推理判断的难度较大，交通事故发生时间较短，时序建模对短时间的异常事件检测的优化效果有限，跨越长时序的历史信息使得聚合的特征对极短的异常事件判断有影响。值得注意的是，虐待这一类异常事件相比基线方法大幅下降。经分析，测试集中虐待类别的视频数量为 2，规模太小从而导致误差被过度放大。

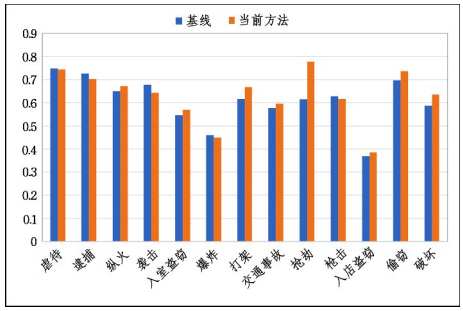

图 7　13 类异常事件的 AUC 结果对比条形图

Fig. 7　AUC of different anomaly detection

下文分别从纵火、抢劫和正常类这 3 类视频中随机选出一例进行可视化展示。

图 8 给出了纵火视频的测试结果。图中阴影区域为异常时间段，其中标注的起始处为刚刚燃起的小火苗，示例分类器没有及时将其判断为异常事件，直到燃烧为明亮的大火苗时才获得了较高的异常置信值。相比基线方法，通过时序建模可以在明火出现后提前做出高响应的预警。

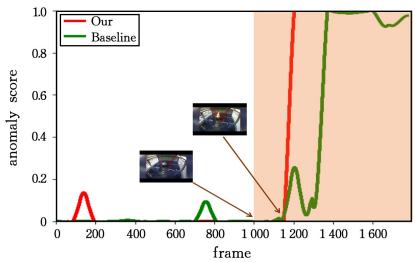

图 8　纵火视频的测试结果

Fig. 8　Example of arson detection

图 9 给出了抢劫视频的检测结果。视频内容是两个抢劫犯分别持枪和刀抢劫一家超市，视频起始处一个持抢的歹徒与超市工作人员交涉，这与正常顾客的购物流程相似，模型出现了漏检。后续视频的发展是一个歹徒持刀洗劫超市物品，另一个歹徒持枪胁迫工作人员交钱，大部分异常事件发生的时序区间都被较好地检测了出来。

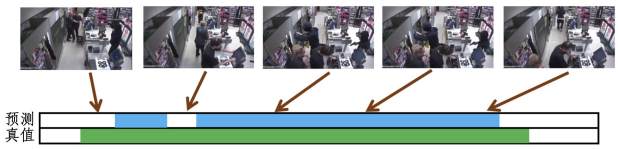

图 9　抢劫视频的测试结果

Fig. 9　Example of robbery detection

图 10 给出了正常视频，记录的是一段街区交通状况的监控视频，没有发生任何异常事件。视频中双向车道车辆往来频繁，其中有一小段发生了低置信度的误检，误检片段是摩托车从侧方超车的画面。

图 10　正常视频的测试结果

Fig. 10　Example of detection on a normal video clip

结束语　本文提出了基于时序建模的弱监督异常事件在线检测算法，该算法以多示例排序网络为基础框架，充分发掘了包与示例间的关系。另外，从视频整体和局部的角度出发展开研究，探索了在线视频异常事件检测的时序模型。最后，在 UCF-Crime 数据集上验证了本文模型的实验效果。

由于视频监控在线检测的特点，模型只能利用历史和当前视频片段的特征，而不能使用未来的视频特征。本文选择 LSTM 和时序卷积对该场景下的视频特征做了时域上下文建模。其中，时序卷积又分为单一时序空洞卷积和金字塔时序空洞卷积，在对当前示例的分类过程中，充分考虑了相邻历史阶段的视频片段。实验结果表明，金字塔时序空洞卷积获得了更好的效果，混合多空洞率的做法充分发掘了视频时域上的多尺度信息。

参 考 文 献

[1] BAI S, KOLTER J Z, KOLTUN V. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling[J]. arXiv: 1803. 01271, 2018.

[2] SULTANI W, CHEN C, SHAH M. Real-world anomaly detection in surveillance videos[J]. arXiv: 1801. 04264, 2018.

[3] BILEN H, VEDALDI A. Weakly supervised deep detection networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 2846-2854.

[4] TANG P, WANG X, BAI X, et al. Multiple instance detection network with online instance classifier refinement[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2843-2851.

[5] LI D, HUANG J, LI Y, et al. Weakly supervised object localization with progressive domain adaptation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 3512-3520.

[6] ZHANG Y, BAI Y, DING M, et al. W2f: A weakly-supervised to fully-supervised framework for object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 928-936.

[7] NGUYEN P, HAN B, LIU T, et al. Weakly supervised action localization by sparse temporal pooling network[C]// 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018.

[8] ZHOU B, KHOSLA A, LAPEDRIZA A, et al. Learning deep features for discriminative localization[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016

[9] PAUL S, ROY S, ROY-CHOWDHURY A K. W-talc: Weakly-supervised temporal activity localization and classification[C]// Proceedings of the European Conference on Computer Vision. 2018: 563-579.

[10] LEE P, UH Y, BYUN H. Background suppression network for weakly-supervised temporal action localization[J]. arXiv: 1911. 09963. 2019.

[11] HASAN M, CHOI J, NEUMANN J, et al. Learning temporal

regularity in video sequences[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 733-742.

[12] LU C, SHI J, JIA J. Abnormal event detection at 150 fps in matlab[C]// Proceedings of the IEEE International Conference on Computer Vision. 2013: 2720-2727.

[13] ZHAO Y, DENG B, SHEN C, et al. Spatio-temporal autoencoder for video anomaly detection [C]// Proceedings of the 2017 ACM on Multimedia Conference. ACM, 2017: 1933-1941.

[14] LIU W, LUO W, LIAN D, et al. Future frame prediction for anomaly detection — a new baseline [J]. arXiv: 1712. 09867, 2017.

[15] DOSOVITSKIY A, FISCHER P, ILG E, et al. Flownet: Learning optical flow with convolutional networks[C]// 2015 IEEE International Conference on Computer Vision (ICCV). 2015.

[16] IONESCU R T, KHAN F S, GEORGESCU M I, et al. Object-centric auto-encoders and dummy anomalies for abnormal event detection in video[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019.

[17] GONG D, LIU L, LE V, et al. Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection[C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV). 2019.

[18] ZHANG J G, QING L Y, MIAO J. Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection[C]// Proceedings of IEEE International Conference on Image Processing. 2019: 4030-4034.

[19] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3d convolutional networks[C]// Proceedings of the IEEE International Conference on Computer Vision. 2015: 4489-4497.

[20] ZHU Y, NEWSAM S. Motion-aware feature for improved video anomaly detection[J]. arXiv: 1907. 10211, 2019.

[21] WANG W, PENG X, QIAO Y, et al. A comprehensive study on temporal modeling for online action detection[J]. arXiv: 2001. 07501, 2020.

[22] OORD A V D, DIELEMAN S, ZEN H, et al. Wavenet: A generative model for raw audio[J]. arXiv: 1609. 03499, 2016.

[23] LI J, ZHANG S, WANG J, et al. Global-local temporal representations for video person reidentification[C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV). 2019.

[24] DUCHI J, HAZAN E, SINGER Y. Adaptive subgradient methods for online learning and stochastic optimization[J]. Journal of Machine Learning Research, 2011, 12(7): 2121-2159.

QING Lai-yun, born in 1974, Ph.D, professor, Ph.D supervisor, is a member of China Computer Federation. Her main research interests include multimedia, computer vision and machine learning.