ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 37,No. 1,Jan. 2022,pp. 94－107
DOI:10. 16337/j. 1004-9037. 2022. 01. 008

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

融合注意力机制的双路径孪生视觉跟踪方法

谢　江[1,2]，朱　艳[1,2]，沈　韬[1,2]，曾　凯[1,2]，刘英莉[1,2]

(1. 昆明理工大学信息工程与自动化学院,昆明 650500;2. 昆明理工大学云南省计算机技术应用重点实验室,昆明 650500)

摘　要：传统基于孪生网络的视觉跟踪方法在训练时是通过从大量视频中提取成对帧并且在线下独立进行训练而成,缺乏对模型特征的更新,并且会忽略背景信息,在背景驳杂等复杂环境下跟踪精度较低。针对上述问题,提出了一种融合注意力机制的双路径孪生网络视觉跟踪算法。该算法主要包括特征提取器部分和特征融合部分。特征提取器部分对残差网络进行改进,设计了一种双路径网络模型;通过结合残差网络对前层特征的复用性和密集连接网络对新特征的提取,将2种网络拼接后用于特征提取;同时采用膨胀卷积代替传统卷积方式,在保持一定感受视野的情况下提高了分辨率。这种双路径特征提取方式可以隐式地更新模型特征,获得更准确的图像特征信息。特征融合部分引入注意力机制,对特征图不同部分分配权重。通道域上筛选出有价值的目标图像信息,增强通道间的相互依赖;空间域上则更加关注局部重要信息,学习更丰富的上下文联系,有效地提高了目标跟踪的精度。为证明该方法的有效性,在OTB100和VOT2016数据集上进行验证,分别使用精确率(Precision)、成功率(Success rate)和平均重叠期望(Expect average overlaprate,EAO)作为评价标准。结果显示,本文算法的精确率、成功率和平均重叠期望分别为0.868、0.641和0.350;相比基准模型分别提高了5.1%、2.0%和0.9%。结果证明本文算法充分利用了不同网络的优点,在保证模型精度的同时,能够较好地适应目标外观的变化,降低相似物的干扰,取得更稳定的跟踪效果。

关键词：目标跟踪;孪生网络;双路径网络;注意力机制;特征融合

中图分类号：TP391　　**文献标志码**:A

Dual-Path Siamese Network Visual Tracking Method with Attention Mechanism

XIE Jiang[1,2], ZHU Yan[1,2], SHEN Tao[1,2], ZENG Kai[1,2], LIU Yingli[1,2]

(1. Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China; 2. Yunnan Key Laboratory of Computer Technologies Application, Kunming University of Science and Technology, Kunming 650500, China)

Abstract: Traditional visual tracking methods based on the Siamese network extract pairs of frames from a large number of videos and train them on the offline independently at the stagey of training. They lack the update of the model features and neglect the background information, so the tracking accuracy is a little bit low in the complex environments such as background clutter. In response to the above problems,

基金项目:国家自然科学基金(61971208, 61671225, 52061020, 61702128);云南省应用基础研究计划重点项目(2018FA034);云南省中青年学术技术带头人后备人才计划(Shen Tao, 2018);云南省万人计划青年拔尖人才计划(沈韬,朱艳,云南省人社厅No. 2018 73);昆明理工大学人才培养计划(KKSY201703016)。
收稿日期:2021-03-26;**修订日期**:2021-06-16

this paper proposes a dual-path Siamese network visual tracking method with the attention mechanism. The method mainly includes the feature extractor part and the feature fusion part. In the feature extractor part, the residual network is improved and a dual-path network model is designed. By combining the reusability of the residual networks to features of the former layer and the extraction of new features from the dense networks, these two networks are spliced for the feature extraction. At the same time, this paper uses the dilated convolution to replace the traditional convolution, which improves the resolution on the condition of maintaining a certain receptive field. This dual-path feature extraction method can implicitly update the model features, so that obtain the more accurate image feature information. Moreover, the attention mechanism is introduced to the feature fusion part, which can distribute the different weights to the different parts of the feature maps. In the channel domain, the method screens the valuable target image information and enhances the interdependence between the channels. In the spatial domain, it also pays more attention to the local important information and learns more rich contextual connections, which effectively improves the accuracy of object tracking. To confirm the effectiveness of the method, some experiments are conducted on the OTB100 and VOT2016 datasets. We use precision, success rate and expect average overlap-rate as the evaluation criterion, and their values are 0.868, 0.641 and 0.350 respectively on the two datasets, which increase by 5.1%, 2.0% and 0.9% compared with those of the benchmark model. Experimental results show that the proposed method makes full use of the advantages of different networks, and while ensuring the accuracy of the model, it can adapt to the deformation of the target well, reduce the interference between the similar objects, and achieve more stable tracking effect.

Key words: object tracking; Siamese network; dual-path network; attention mechanism; feature fusion

引　　言

视觉目标跟踪是指在不断变化的视频序列中自动定位特定目标,是各种计算机视觉任务的基本组成部分,在视频监控[1]、计算机交互[2]和扩增现实[3]等领域有着广泛的应用,逐渐成为计算机视觉领域研究热点之一。而由于现实环境的复杂和多样性,当目标处于遮挡、形变、快速运动或相似物背景干扰等一系列具有挑战性的场景中时,如何准确且快速有效地检测和定位目标是亟待解决的问题。

近年来,从SINT[4]提出开始,采用相似度对比策略、结合深度学习从而实现跟踪的方法逐渐被关注。该方法开创性地将视觉目标跟踪问题转化为patch块匹配问题,通过计算目标模板和候选区域的相似度,运用两者之间的互相关操作完成跟踪。其后SiamFC[5]沿用patch块匹配方法,设计了一种端到端的跟踪网络,使用更大的数据集进行训练,在Titan xp上达到了58帧/s。对比同期基于相关滤波的算法,基于孪生网络的跟踪器因为其简易的框架和较高的跟踪速度受到关注。CFNet[6]将相关滤波操作融入神经网络,将其作为单一的网络层嵌入到网络之中,进一步提升了网络的性能。DSiam[7]在SiamFC框架上加入目标外观变换转换层和背景抑制变换层来提升网络的判别能力,增强了模型的在线更新能力。SA-Siam[8]使用双网络训练,分别获取网络的语义特征和外观特征,之后进行互相关操作并将整个网络整合到一起,有效地增强了目标的识别和定位能力。SiamRPN[9]则在孪生网络之后引入了区域建议网络(Region proposal network, RPN),将相似度计算问题转化为分类和回归问题;采用分类分支和回归操作取代原本的全连接层获得网络模型,在保证跟踪速度的同时大大提升了模型性能。DaSiamRPN[10]从数据输入角度出发,平衡训练集正负样本不均衡问题和样本的丰富性问题,使得跟踪模型的泛化能力得到较大提升。CIR[11]在SiamRPN的基础框架下,加入Crop操作,通过堆叠CIR模块将

Padding造成的影响裁去,进而构建了深层网络,成功解决了随着网络层数加深跟踪效果反而降低的问题。同样对于网络中Padding操作引起的网络更加关注图像中心的问题,SiamRPN++[12]从采样策略角度出发,对图像进行移位操作,将拥有较大权重的图像中心进行均匀采样;同时级联多个RPN网络,抽取深度网络的多个特征层分别进行分类和回归,有效地解决了相似目标的误检问题。虽然这些跟踪方法很好地平衡了检测精度和跟踪速度,获得了不错的结果,但仍存在以下问题需要解决。

(1)在特征提取阶段,Siamese跟踪方法难以区分前景和非语义背景,非语义背景通常被认为是干扰因素;当背景杂乱或者附近存在相似物时,目标特征提取不准确,难以保证跟踪性能。

(2)大多数孪生跟踪器不能更新模型[4-5,11-12]。由于这些跟踪器模型固定且结构简单,从而失去了在线更新外观模型的能力,难以解释在跟踪场景中目标剧烈的外观变化。同时这些孪生跟踪器采用了局部搜索策略,当目标处于完全遮挡或视野之外的场景中时,跟踪器会丢失目标。

针对这两点问题,本文提出了融合注意力机制的双路径孪生视觉跟踪方法,主要贡献如下:

(1)提出双路径特征提取方法,充分利用残差网络(Residual network, ResNet)对旧特征的复用性和密集连接网络(Dense network, DenseNet)对新特征的探索,在保证目标特征表示的同时增强了跟踪的鲁棒性。同时网络中卷积核采用膨胀卷积,在保持一定感受视野的情况下提高分辨率。

(2)引入基于注意力机制的特征融合,通过注意力机制融合不同尺度、纹理、位置等信息,增强了目标的特征表示能力,进一步提高了跟踪精度。

1 相关理论

1.1 基于孪生网络的特征提取子网络

1.1.1 AlexNet

AlexNet创新性地采用了非线性激活函数ReLU,取代了传统算法中支持向量机(Support vector machines, SVM)等人工设计的特征,成功解决了Sigmoid在网络较深时的梯度弥散问题。提出了局部相应归一化(Local response normalization, LRN),将其加在激活函数和池化函数之后,加大特征响应较大的区域,增强了模型的泛化能力;同时使用数据增广和Dropout防止过拟合。AlexNet网络框架如图1所示。

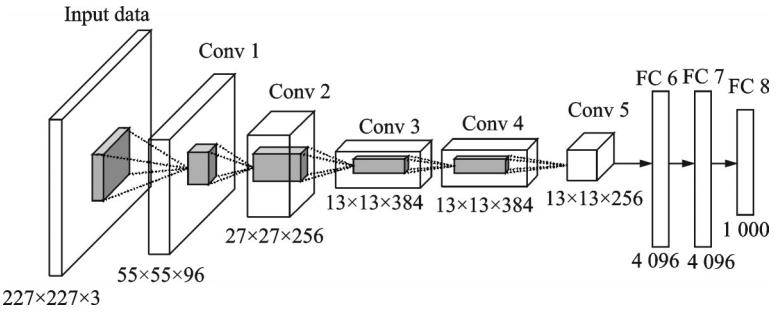

图1 AlexNet网络结构图

Fig.1 Structure diagram of AlexNet

1.1.2 ResNet

随着深度学习的发展,网络层数的加深,模型变得越来越复杂,从而导致随机梯度下降法(Stochastic gradient descent, SGD)的优化变得更加困难,出现“退化”现象,模型效果反而降低。对此何凯明等[13]提出ResNet网络,引入了能够跳过一层或多层的“shortcut connection”,如图2所示。图中ResNet包含2种

映射:Identity映射和Residual映射,最后输出为$y=F(x)+x$,其中x表示Identity映射,$F(x)$表示Residual映射。在神经网络计算梯度时,由本身经过权重层的Identity映射加上没有经过权重层衰减的跳层连接Residual映射,有效避免了梯度消失问题。

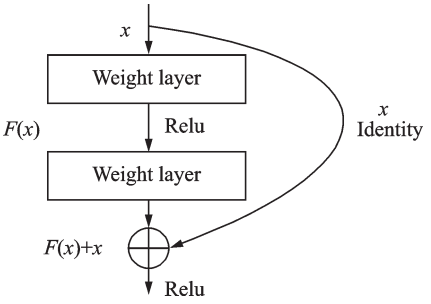

图2　跳层连接示意图

Fig.2　Shortcut connection schematic

1.1.3　DenseNet

DenseNet网络在ResNet基础上采用更密集的连接方式,以前馈的方式将每一层相互连接,具体网络结构如图3所示。从图3中可以发现,DenseNet网络将所有网络层连接在一起,使各层信息交互达到最大。对于每一层而言,前面所有层的特征映射都用作输入,而它自己的特征映射也会用作后面所有层的输入。该方法有效缓解了梯度消失问题,减少了信息在网络层传播的流失,增强了对前层信息的重用和对新特征的提取。

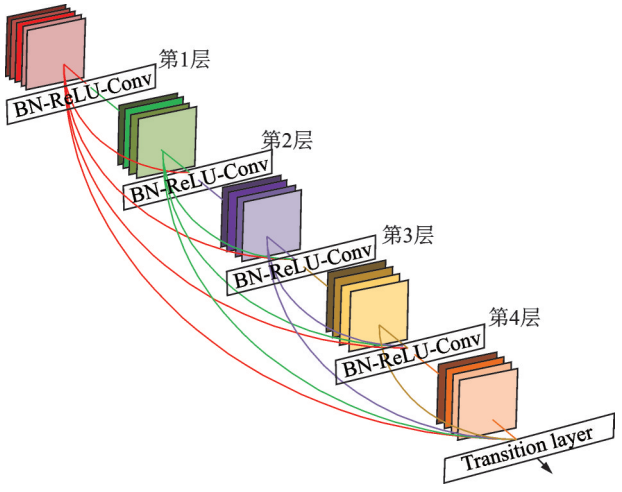

图3　DenseNet网络结构图

Fig.3　Structure diagram of DenseNet

1.2　基于孪生网络的特征提取子网络

注意力机制分为软注意力和硬注意力。其中软注意力是一种确定性注意力,具有可微性,在网络中经过前向传播和反向传播得到注意力的权重,从而给予所关注的信息更多的权重,实现模型效果的提升。软注意力主要分为通道注意力和空间注意力。

1.2.1　通道注意力机制

通道注意力类似信号变换,在图像经由网络变换后输出多通道特征。对于每个通道而言,其对目标关键特征的贡献不同,为此通过给予不同通道不同权重,特征相关度越高则给予越高的权重,从而有效提高特征提取的效果。

1.2.2　空间注意力机制

在整个特征空间中,目标仅占整个图像的一小部分,为此需要找出目标关注的区域。空间注意力可以将原始图像信息转换到空间域,并能够保留其中的关键信息;通过将原始图片中的空间信息变换到另一个空间中,训练出的空间注意力能够保留图像中的关键信息,从而实现对上层关键信息的识别。网络结构图如图4所示。

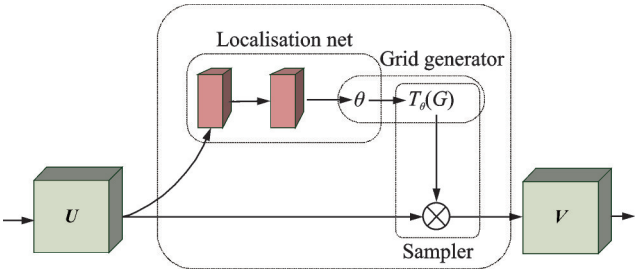

图4 空间注意力结构图

Fig.4 Structure diagram of spatial attention

1.3 多层特征融合

图像处理过程中需要对特征图进行采样，特征图的大小不可避免地会发生改变，并且越低层的特征其分辨率越高，位置信息和细节信息越完善；反之，越高层的特征，则语义信息更强。因此如何有效地融合多层特征、取长补短，是改善模型性能的关键。经典方法有特征金字塔[14](Feature Pyramid network, FPN)方法和自注意力方法等。

2 网络设计

2.1 整体网络框架

本文网络的整体框架采用孪生网络结构，如图5所示。本文算法基本流程为：首先将目标和搜索模板送入特征提取模块，该模块由双路径孪生网络构成；之后对特征进行处理，采用设计的结合注意力机制的特征融合模块，由注意力机制和特征融合方法构成。

其中目标模板和搜索模板所采用双路径网络特征提取结构如图6所示。双路径网络包含2部分：第1部分是紫色矩形框，表示ResNet，通过残差块跳层连接网络下一层，但保持通道数不变；第2部分是蓝色多边形框，通过对输入做1×1、3×3和1×1卷积，再和该层输出做通道合并，即Concat操作得到

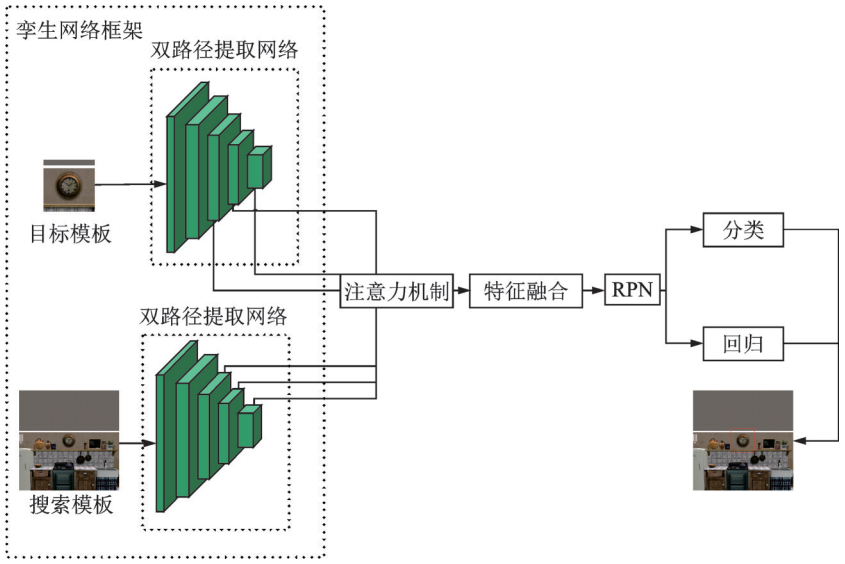

图5 整体网络框架图

Fig.5 Frame diagram of whole network

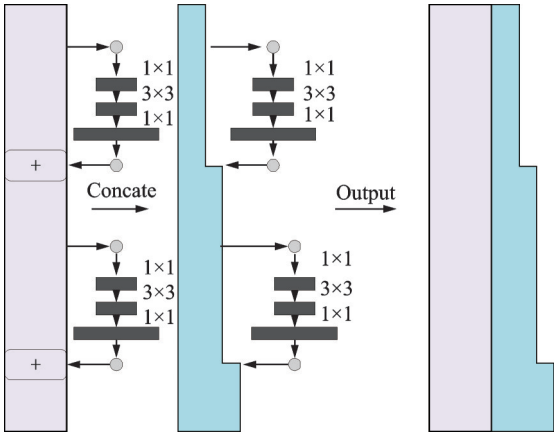

图6　双路径网络结构图

Fig.6　Structure diagram of DualpathNet

Dense部分。

本文提出的方法不仅获得更精细的特征,并且随着目标外观的变化更新了特征信息。同时学习人类视觉机制设计了注意力模块用以提高跟踪的准确率。最后将卷积层的多层特征进行融合,将采集到的不同感受野的特征融合在一起,获得更精确的目标特征表示和位置信息。

2.2　膨胀卷积

为了在保持一定感受视野的同时提高分辨率,本文算法在双路径网络中使用了膨胀卷积。普通卷积的表达式为

$$O(x,y)\cdot H(x,y)=\sum_{i=0}^{w}\sum_{j=0}^{h}H(x,y)\cdot O(x-i,y-j) \tag{1}$$

式中:$O(x,y)$为原始图像在点(x,y)处的像素值;$H(x,y)$是与其相乘的卷积核,大小为$w\times h$。

膨胀卷积计算为

$$O(x,y)\cdot H'(x,y)=\sum_{i=0}^{w}\sum_{j=0}^{h}H'(x,y)\cdot O(x-l\times i,y-l\times j) \tag{2}$$

式中:l为膨胀因子;$H'(x,y)$为膨胀卷积核。

从式(2)可以看出,膨胀卷积实质上就是对卷积核进行了0填充,这样做可以在增加卷积核感受视野的同时保留原始的像素信息,增大了分辨率。若卷积核的尺寸为k,膨胀率为l,则膨胀卷积的实际有效尺寸为$k+(k-1)\times(l-1)$。与相同大小的普通卷积相比,膨胀卷积不仅扩大了感受视野,还保持了与普通卷积相同的分辨率,如图7所示。

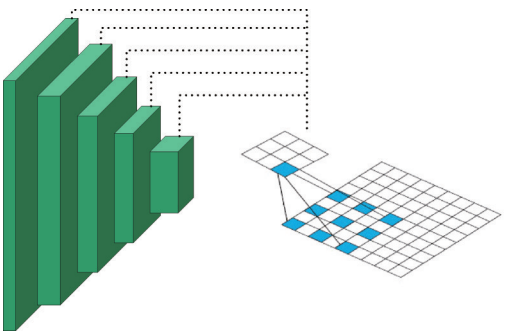

图7　膨胀卷积结构图

Fig.7　Structure diagram of dilated convolution

2.3　注意力机制

为了弥补下采样造成的细节丢失,更好地指导模型的训练,本文采用一种注意力机制,通过对特征图进行加权处理从而增强目标特征并且抑制干扰背景。改进的注意力机制主要由通道注意力和空间注意力组成,SE(Squeeze-and-exciation)模块如图8所示。从图8可以看出,该模块通过将卷积层输出结果送入一个通道注意力

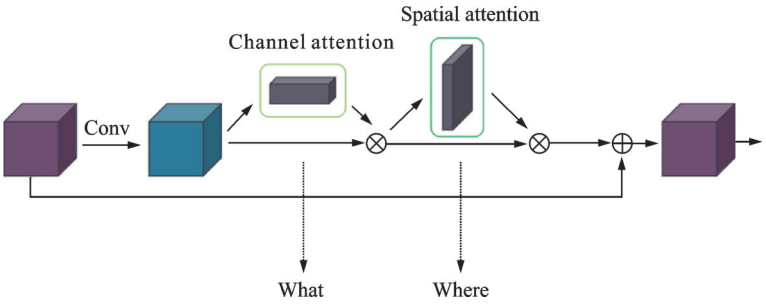

图8 SE模块结构图

Fig.8 Structure diagram of SE block

模块,得到加权结果之后;再送入一个空间注意力模块,最终进行加权得到结果。

为研究注意力机制对特征关注区域的影响,通过在骨干网络后接入注意力模块,给予目标中心区域更大的权重,以提高模型对目标检测的精度。在COCO和ILSVRC_DET数据集上,以双路径网络作为特征提取框架进行训练,其可视化CAM热力图如图9(a)所示,经过注意力模块给予不同特征信息不同权重后的CAM热力图如图9(b)所示。从两个图对比可知,添加注意力模块后,目标关注区域更集中在目标中心,减少了对非目标区域特征的权重,从而提升了特征提取效果。

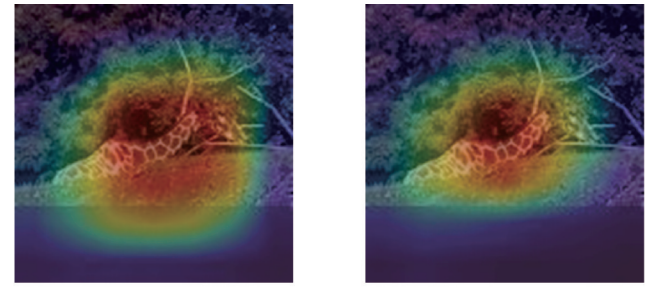

图9 可视化CAM热力图

Fig.9 Visualized CAM heat map

2.4 特征融合模块

在跟踪过程中,由于视角等一系列问题,目标大小会发生变化。当目标较大时,大感受视野获取的特征比较重要;而当目标较小时,较大的感受视野会采集过多的周边信息,从而造成误差。传统的特征融合方法一般是级联或者直接相加,这种方法简单但是没有考虑不同特征图的感受视野不同,忽略了特征之间的特异性。因此,本文设计特征融合模块对不同感受视野下的特征图分配不同权重,实现了高效的特征融合,如图10所示。

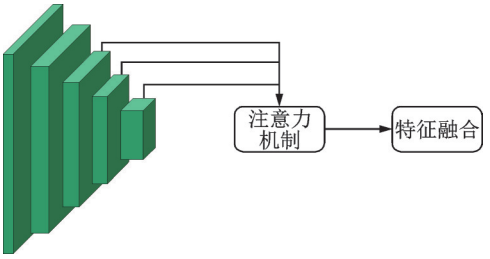

图10 特征融合模块

Fig.10 Feature fusion block

在图像特征提取之后,模型提取Conv 3、Conv 4、Conv 5的特征信息,将不同层次的特征表达先送入注意力模块,获得重要通道和目标所在空间的信息;再送入特征融合模块,融合多层卷积后输出的特征,在一定程度上提高了特征的多样性,从而提高了模型的性能。通过特征融合模块,为不同感受视野下的特征图分配权重,使得不同感受

视野下的特征特异性得到了体现,使特征更好地进行融合。

2.5　目标分类和回归模块

在特征提取、特征融合之后,接入RPN模块,对目标进行分类和位置回归。其网络结构图如图11所示。

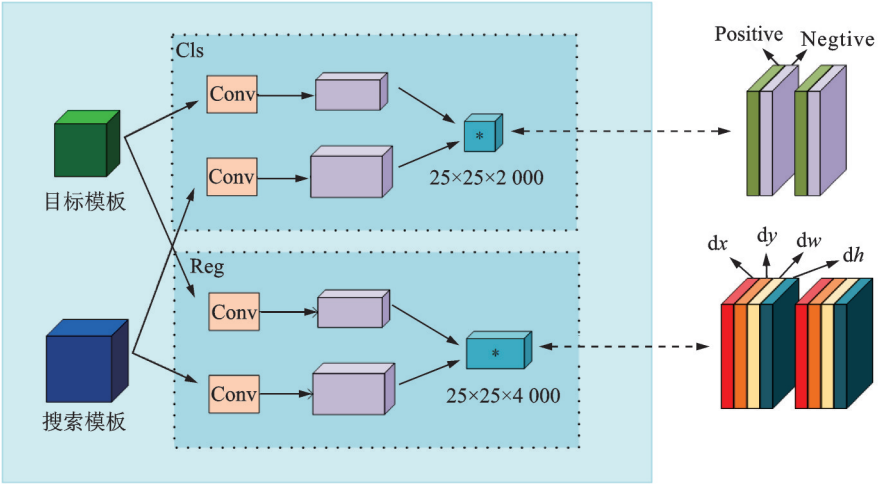

图11　目标分类和回归模块

Fig.11　Target classification and regression Block

从图11可见,在特征融合后,将目标模板和搜索模板经过一系列特征操作之后送入RPN模块。该模块将目标模板卷积后所得特征图作为卷积核在搜索模板卷积所得的特征图上进行互相关操作,对应于图中的Cls分支和Reg分支;其大小分别为卷积核大小的2 000倍和4 000倍。

3　实验结果与分析

本文实验训练数据集采用单目标跟踪领域常用的公共数据集COCO和ILSVRC_DET,测试数据集采用OTB100和VOT2016。COCO数据集中图像主要从复杂日常场景中截取,图像包括91类目标,328 000个影像和2 500 000个标签。ILSVRC_DET数据集是ILSVRC竞赛中用于目标检测任务的数据集,包含200个物体类和数万张照片。2个测试集的示例和本文算法在2个数据集上的结果展示如图12所示。图12包含6组图像:绿色框表示目标标签;黄色框表示本文算法的跟踪结果。在数据集上测试时,仅读取目标首帧的位置标签,然后通过矩形框在后续帧对目标进行框定,从而实现跟踪。对于一般视频序列而言,只需在任意帧中手动框选出目标,再通过加载离线训练完成的模型,则可实现后续帧对目标的自动框定、跟踪。

3.1　评价指标

OTB100中使用精确率曲线图(Precision plot)和成功率曲线图(Success plot)作为评价标准。精确率曲线图描述的是跟踪算法预测的目标框和标注的目标框之间中心位置的误差(Center location error,CLE),表达式为

$$\mathrm{GLE}=\sqrt{(X_{\mathrm{T}}-X_{\mathrm{G}})^2+(Y_{\mathrm{T}}-Y_{\mathrm{G}})^2} \tag{3}$$

式中:$(X_{\mathrm{T}}, Y_{\mathrm{T}})$为预测框中心坐标;$(X_{\mathrm{G}}, Y_{\mathrm{G}})$为真实框中心坐标。

精确率采用在精确曲线图中位置错误阈值为20时所对应的精确度值。成功率曲线则是指重叠率

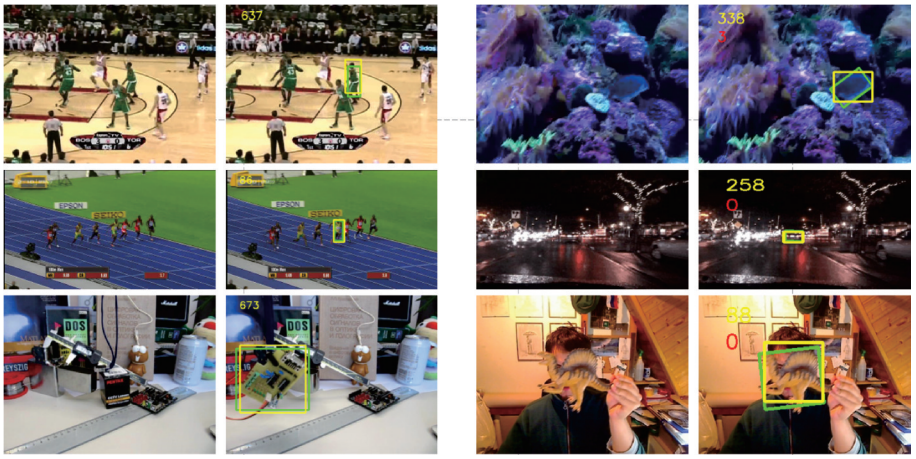

图12 样例图

Fig.12 Sample images

与重叠阈值之间的曲线，其值以该曲线图的线下面积作为成功率的依据。重叠率得分(Overlap score，OS)计算公式为

$$\mathrm{OS}=\frac{A_{\mathrm{T}}\cap A_{\mathrm{G}}}{A_{\mathrm{T}}\cup A_{\mathrm{G}}} \tag{4}$$

式中：A_{T}，A_{G}分别表示预测框与真实框的区域；∩表示两个框面积的交集；∪表示两个框面积的并集。

VOT2016中使用精确度(Accuracy)、鲁棒性(Robustness)和期望平均重叠率(Expect average overlaprate，EAO)作为评价标准。精确度由跟踪算法预测的目标框和标注的真实框的重叠率得到，计算公式为

$$\mathrm{Accuracy}=\frac{1}{N_{\mathrm{valid}}}\sum_{t=1}^{N_{\mathrm{valid}}}\mathrm{OS}_t \tag{5}$$

式中：OS_t表示第t帧的重叠率；N_{valid}表示视频序列的长度。OS_t的计算公式如下

$$\mathrm{OS}_t=\frac{1}{N_{\mathrm{rep}}}\sum_{k=1}^{N_{\mathrm{rep}}}\frac{A_t^k\cap A_{\mathrm{gt}}^k}{A_t^k\cup A_{\mathrm{gt}}^k} \tag{6}$$

式中：A_t^k、A_{gt}^k分别表示第t帧第k次重复得到的预测框与真实框面积；N_{rep}为重复的次数。

鲁棒性则用来测试跟踪算法的稳定性，数值越大，表示稳定性越差，计算公式为

$$\mathrm{Robustness}=\frac{1}{N_{\mathrm{rep}}}\sum_{k=1}^{N_{\mathrm{rep}}}F(k) \tag{7}$$

式中$F(k)$为跟踪算法在跟踪的第k次重复中的失败次数。

期望平均重叠率则通过精确度和鲁棒性计算得到。将数据集中所有序列按长度N_{s}分类，计算长度为N_{s}的视频序列每一帧一次性的精确性，然后计算每个长度N_{s}为视频序列的精确性，该长度N_{s}的EAO值为

$$\mathrm{EAO}_{N_{\mathrm{s}}}=\frac{1}{N_{\mathrm{s}}}\sum_{v=1}^{N_{\mathrm{s}}}\mathrm{Accuracy} \tag{8}$$

最后对不同长度的EAO值求平均，设序列长度范围为[$N_{\mathrm{min}}-N_{\mathrm{max}}$]，则计算公式为

$$\mathrm{EAO}=\frac{1}{N_{\max}-N_{\min}}\sum_{N_s=N_{\min}}^{N_{\max}}\mathrm{EAO}_{N_s} \tag{9}$$

EAO用以评测跟踪算法的综合性能。

3.2　实验结果分析

实验采用环境为Pytorch1.1.0,CUDA版本为10.2,计算机配置为Intel(R) Core(TM) i5-9400F CPU @ 2.90 GHZ,内存16.0 GB,显卡为NVIDIA GeForce RTX 2060。

对输入数据进行预处理,将数据集图像进行填充、裁剪、缩放,最后得到的搜索图像大小为255像素×255像素,模板图像大小为127像素×127像素。使用双路径网络作为基础网络,加载在ImageNet上预训练的权重。网络初始学习率为0.005,衰减率为0.000 1,batch size设置为16,每轮训练600 000次,总共训练20个epoch。

本文基准模型为SiameseRPN++,在此基础上将特征提取子网络ResNet换成双路径网络(DualpathNet),卷积核采用膨胀卷积。为验证方法的有效性,在保持其他网络架构不变的情况下进行实验。结果显示:本文方法的精确度为83.9%、跟踪成功率为63.1%,优于原基准模型。以上实验均在OTB100上进行,具体实验结果如表1所示。由表1可见,采用双路径网络后,目标的跟踪精确度指标和成功率指标分别提升了2.2%和1.0%,证明了该方法的有效性。其中,在低分辨率和超出视野情况下,跟踪精度提升最多,其中精确度分别提高了10%和8.2%;在超出视野和背景驳杂情况下,跟踪成功率提升最多,其成功率分别提高了6.4%和3.4%。具体实验结果如图13,14所示,其中图13中图注方括号中数值表示位置错误阈值为20时的精确度,图14中图注方括号中数值表示跟踪算法成功率曲线下面积的大小,下文图中图注数值含义相同。

表1　不同backbone在OTB100数据集上实验结果对比

Table 1　Experimental results of different backbones on OTB100 dataset

模型	成功率	精确率	速率/(帧·s^{-1})
Baseline(ResNet)	0.621	0.817	**29.3**
Baseline(DualpathNet)	**0.631**	**0.839**	28.9

注:加粗字体表示最优结果

同时,为了验证注意力机制的有效性,在基准网络基础上,对网络添加注意力模块和未添加注意力模块进行了实验。实验在OTB100上进行,实验证明,添加了注意力模块后网络性能得到了提升,实验

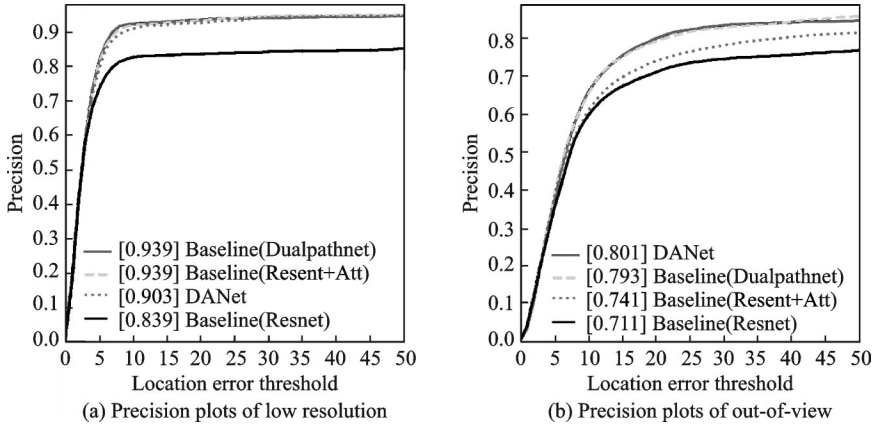

图13　低分辨率和超出视野情况下精确度结果图

Fig.13　Results of precision on low resolution and out of view

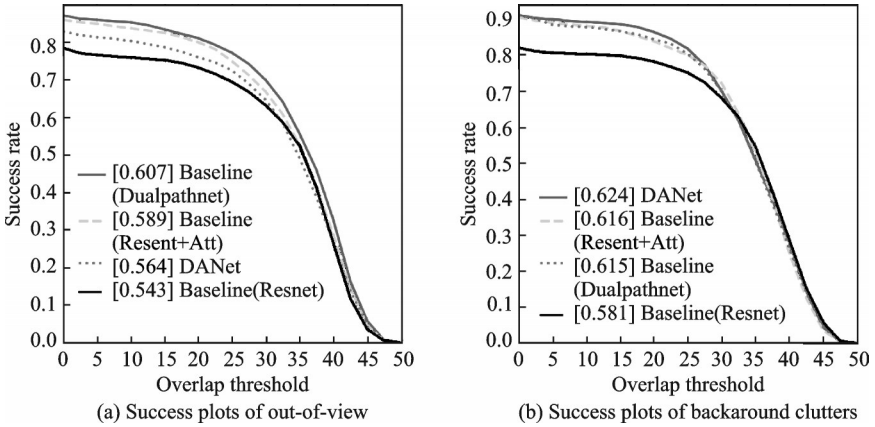

图14 超出视野和背景驳杂情况下成功率结果图

Fig.14 Results of success rate on out of view and background clutters

结果如表2所示。由表2可见,在添加注意力之后,精确度指标提升了2.4%,成功率提升了0.5%,速率下降0.2帧/s。其中,在背景驳杂和形变情况下,精确度分别提高了5.4%和5.1%;在低分辨率和形变情况下,成功率提高了1.2%和2.6%。具体实验结果如图15和图16所示。

表2 OTB100上未添加和添加注意力结果对比

Table 2 Experimental results with and without adding attention on OTB100 dataset

模型	成功率	精确率	速率/(帧·s^{-1})
Baseline(Resnet)	0.621	0.817	**29.3**
Baseline(Resnet+Att)	**0.626**	**0.841**	29.1

注:加粗字体表示最优结果。

最后,在双路径网络基础上,对网络添加注意力模块进行了实验。实验在OTB100上进行,结果证明,本文提出的融合注意力机制的双路径孪生视觉跟踪方法有效地提升了跟踪性能,实验结果如表3所示。由表3可见,本文提出的融合注意力机制的双路径孪生视觉跟踪方法有效地提升了跟踪效果。在本文实验条件下,成功率提升了2%,准确率提升了5.1%,同时在实时性性能上,仅比基准模型降低0.7帧,达到28.6帧/s,满足大部分场景下实时性要求。其综合性能指标如图17所示。

上述结果显示,采用膨胀卷积的双路径网络在低分辨率和超出视野情况下精确度提高显著,在超

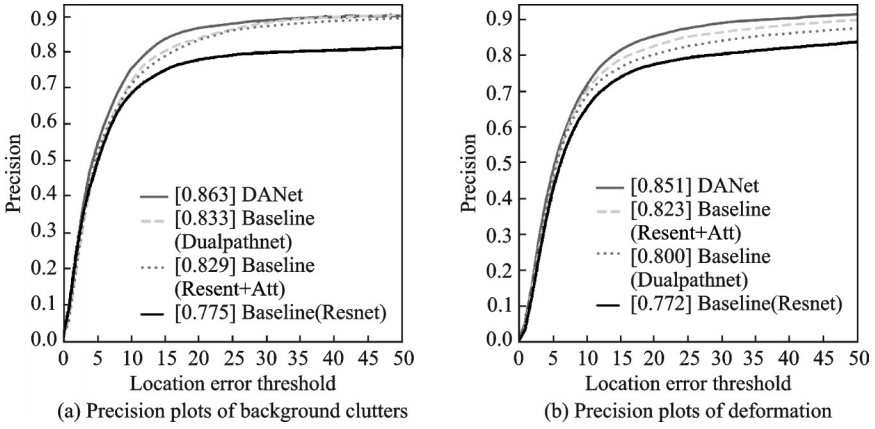

图15 背景驳杂和形变情况下精确度结果图

Fig.15 Results of precision on backround clutters and deformation

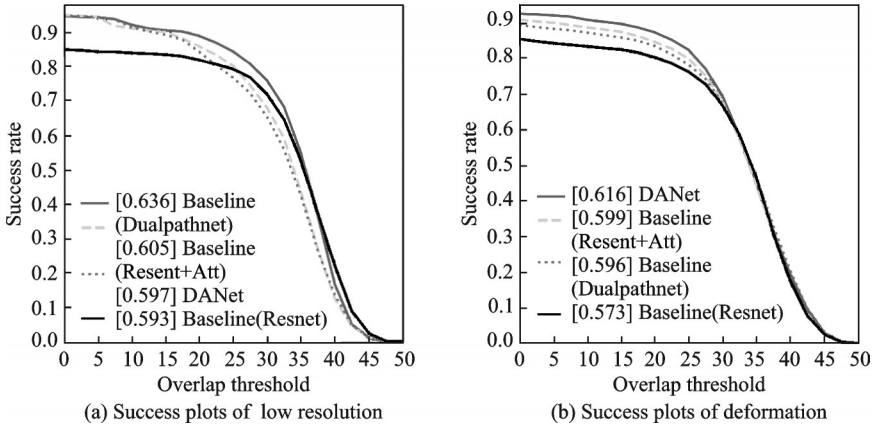

图16　低分辨率和形变情况下成功率结果图

Fig.16　Results of success rate on low resolution and deformation

出视野和背景驳杂情况下成功率提高显著;结合注意力机制的特征融合模块,在背景驳杂和形变情况下精确度提高显著,在低分辨率和形变情况下成功率提高显著。由此得出结论,双路径网络提取特征通过隐形更新模型,获取了更新、更精细的特征;在目标短暂消失和特征模糊的情况下效果显著。结合注意力机制的特征融合模块则结合不同层尺度、纹理、位置信息,有效提高了目标在信息残缺的情况下的跟踪效果。

表3　OTB100上完整消融实验结果

Table 3　Diagram of complete ablation experiment result on OTB100 Dataset

模型	成功率	精确率	速度/(帧·s^{-1})
Baseline(ResNet)	0.621	0.817	**29.3**
Baseline(DualpathNet)	0.631	0.839	28.9
Baseline(Resnet+Att)	0.626	0.841	29.1
DANet	**0.641**	**0.868**	28.6

注:加粗字体表示最优结果。

为了证明本文算法对比其他算法的优势,在相同条件下在OTB100和VOT2016上进行了试验。结果如表4所示。从表4可以看出,在OTB100上本文算法精确度和成功率指标达到了0.868和0.641;在VOT2016上准确度、鲁棒性和EAO分别达到了0.608、0.303和0.350,结果证明本文算法的各项指标都较优。

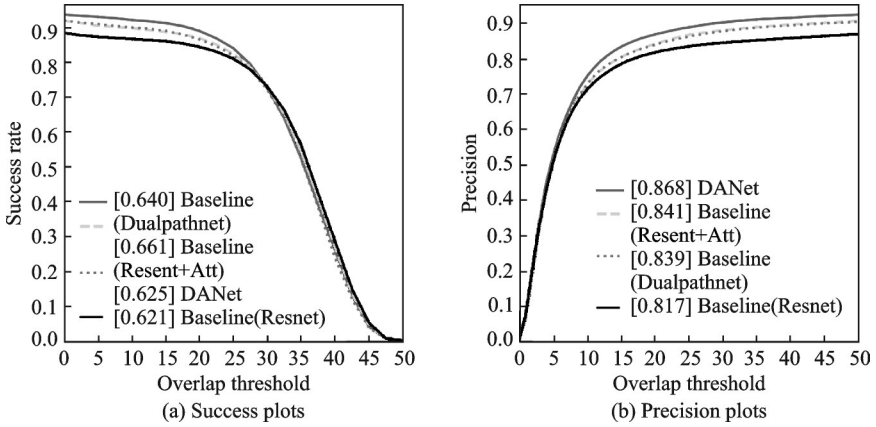

图17　OTB100数据集综合性能指标

Fig.17　Integrated performance index on OTB100 dataset

表4 不同算法在OTB100和VOT2016上的结果对比

Table 4 Comparison of results of different algorithms on OTB100 and VOT2016 datasets

方法	OTB100		VOT2016		
	成功率	精确率	准确率	鲁棒性	EAO
SiamFC	0.580	0.770	0.530	0.460	0.235
MDNet			0.540	0.340	0.257
SiamRPN	0.590	0.800	0.560	**0.260**	0.334
CFNet	0.590	0.780			
Staple	0.580	0.780	0.54	0.38	0.30
Baseline	0.621	0.817	0.579	0.275	0.341
Ours	**0.641**	**0.868**	**0.608**	0.303	**0.350**

注:加粗字体表示最优结果。

4 结束语

本文提出了一种融合注意力机制的双路径孪生视觉跟踪方法,通过双路径网络进行特征提取,卷积核采用膨胀卷积,结合残差网络对特征的复用性和密集网络对特征的重用性,有效地提高了跟踪的精确度。同时在特征提取网络层之间添加注意力机制,使得网络自适应学习权重,并将不同感受野的特征进行融合,有效提高了模型性能。将本文算法在OTB100和VOT2016上进行测试,并与多种算法比较,本文算法取得了较优的结果。但更新模型在适应目标外观变化的同时,丢失了之前学习的特征,导致跟踪过程中鲁棒性不足。下一步将针对序列中模型特征变化角度,改进模型更新策略,探究更新的间隔时间和时间段内的模型权重方法,增强模型的稳定性表达,进一步提升跟踪性能指标。

参考文献

[1] XING Junliang, AI Haizhou, LAO Shihong. Multiple human tracking based on multi-view upper-body detection and discriminative learning [C]// Proceedings of the 20th International Conference on Pattern Recognition. Istanbul, Turkey: IEEE, 2010: 23-26.

[2] ZUO Wangmeng, WU Xiaohe, LIN Liang, et al. Learning support correlation filters for visual tracking [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(5): 1158-1172.

[3] ZHANG Guangcong, VELA P A. Good features to track for visual SLAM [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, MA, USA: IEEE, 2015: 1373-1382.

[4] TAO Ran, GAVVES E, SMEULDERS A W M. Siamese instance search for tracking [C]// Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 2016: 1420-1429.

[5] BERTINETTO L, VALMADRE J, HENRIQUES J F, et al. Fully-convolutional siamese networks for object tracking [C]// Proceedings of the European Conference on Computer Vision Workshop. Amsterdam, Netherlands: Springer, 2016: 850-865.

[6] VALMADRE J, BERTINETTO L, HENRIQUES J F, et al. End-to-end representation learning for correlation filter based tracking [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE, 2017: 5000-5008.

[7] GUO Qing, WEI Feng, ZHOU Ce, et al. Learning dynamic siamese network for visual object tracking [C]// Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017: 1781-1789.

[8] HE Anfeng, CHONG Luo, TIAN Xinmei, et al. A twofold siamese network for real-time object tracking [C]// Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018: 4834-4843.

[9] LI Bo, YAN Junjie, WU Wei, et al. High performance visual tracking with siamese region proposal network [C]//

Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA: IEEE, 2018: 8971-8980.

[10] ZHU Zheng, WANG Qiang, LI Bo, et al. Distractor-aware siamese networks for visual object tracking [C]// Proceedings of the European Conference on Computer Vision Workshop. Munich, Germany: Springer, 2018: 103-119.

[11] ZHANG Zhipeng, PENG Houwen. Deeper and wider siamese networks for real-time visual tracking [C]// Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA: IEEE, 2019: 4586-4595.

[12] LI Bo, WU Wei, WANG Qiang, et al. SiamRPN++: Evolution of siamese visual tracking with very deep networks [C]// Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, CA, USA: IEEE, 2019: 4277-4286.

[13] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition [C]// Proceedings of the 2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Las Vegas, NV, USA: IEEE, 2016: 770-778.

[14] LIN T Y, DOLLAR P, GIRSHICK R, et al. Feature pyramid networks for object detection [C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA: IEEE, 2017: 936-944.

作者简介:

谢江(1995-),男,硕士研究生,研究方向:计算机视觉和图像处理,E-mail:940351559@qq.com。

朱艳(1979-),**通信作者**,女,博士,教授,硕士生导师,研究方向:智能检测和人工智能,E-mail:zhuyan@kust.edu.cn。

沈韬(1984-),男,教授,研究方向:太赫兹材料无损检测、辅助驾驶和人工智能,E-mail:shentao@kust.edu.cn。

曾凯(1985-),男,副教授,研究方向:粒计算和分布式计算,E-mail:zengkailink@sina.com。

刘英莉(1978-),女,讲师,研究方向:材料文本挖掘和深度学习,E-mail:lyl2002@126.com。

(编辑:张黄群)