DOI:10.13878/j.cnki.jnuist.20231113002

邬心怡[1]　邓志良[1]　刘云平[1]　董娟[2]　李嘉琦[1]

基于交叉注意力机制的多特征行人重识别

摘要

针对现有的行人重识别方法难以避免环境噪声导致的特征提取不精确、易被误认为行人特征等问题,提出一种基于动态卷积与注意力机制的行人多特征融合分支网络.首先,由于拍摄时存在光照变化、人体姿势调整以及物体遮挡等不确定因素,提出使用动态卷积替换ResNet50中的静态卷积得到具有更强鲁棒性的Dy-ResNet50模型;其次,考虑到拍摄行人图片的视角有较大差异且存在行人被物体遮挡的情况,提出将自注意力机制与交叉注意力机制嵌入骨干网络;最后,将交叉熵损失函数和难样本三元损失函数共同作为模型损失函数,在DukeMTMC-ReID、Market-1501和MSMT17公开数据集上进行实验,并与主流网络模型进行比较.结果表明:在3个公开数据集上,本文所提模型的Rank-1(第一次命中)与mAP(平均精度均值)相比当前主流模型均有所提升,具有较高的识别准确率.

关键词

行人重识别;动态卷积;自注意力机制;交叉注意力机制

中图分类号 TP391.4

文献标志码 A

收稿日期 2023-11-13

资助项目 国家自然科学基金(51875293);国家重点研发计划(2018YFC1405703)

作者简介

邬心怡,女,硕士生,研究方向为计算机视觉、图像处理.2661491003@qq.com

邓志良(通信作者),博士,教授,研究方向为智能控制.002858@nuist.edu.cn

1 南京信息工程大学 自动化学院,南京,210044

2 南京信息工程大学 电子与信息工程学院,南京,210044

0 引言

行人重识别(person Re-identification,Re-ID)[1]是计算机视觉领域的一项任务,目的是在多个摄像头之间识别并匹配出同一个行人,该技术的研究对于智能监控、图像检索、刑事侦查等领域具有重要的现实意义.然而,由于光照条件、行人姿态、拍摄背景等不确定因素,同一行人在不同条件下表现出来的特征差别很大,且容易出现图像模糊不清或被遮挡的情况,使得行人重识别任务面临很大挑战.因此,如何有效地提取出具有较强可辨识性、强鲁棒性的特征是当前行人重识别领域的一个热点问题.

行人重识别任务近年来得到了广泛研究.随着深度学习的发展,学者们通过对深度特征图进行分块,使网络关注更小的区域,从而提取行人局部细节信息[2-3],但此类方法过度注重局部而忽略全局信息,模型识别准确率不高;还有一些方法通过改进距离度量[4-5],比较行人图片,缩短同一身份行人图像的特征距离,并使无关特征远离本身份簇,但在实际应用中由于拍摄场景存在遮挡和背景冗余等问题,此类方法无法正确提取行人有效特征.

尽管现有大多数Re-ID模型已具备较好的识别能力,但是同一个行人在不同摄像机拍摄下,正面与侧面存在较大的视角差异导致对于输入图片间的特征交互还不够充分.为了弥补现有方法的缺陷,提高网络模型的鲁棒性,本文以ResNet50作为骨干网络,设计了一种基于动态卷积(Dynamic Convolution)[6]与注意力机制(Attention Mechanism)的多特征融合分支网络.首先,由于拍摄时的光照、人体姿势以及拍摄视角等不确定因素,识别特征难度较大,本文提出使用动态卷积根据输入的不同灵活调整卷积核权重,以便高效提取行人有效特征.其次,考虑到拍摄行人图片的视角有较大差异且存在行人被物体遮挡的情况,为了充分获取行人特征及输入图片间的特征交互信息,本文提出将自注意力机制[7]与交叉注意力机制[8]嵌入骨干网络,使计算机能更好地模拟人类视觉识别图像特征.最后,在改进距离度量方面,本文采用交叉熵损失函数(Cross Entropy Loss)[9-10]和难样本三元损失函数(TriHard Loss)[11-12]共同作用模型,将提取到的行人全局特征与局部特征融合后进行分类与匹配.与现有主流网络模型的对比分析结果表明,本文模型具有较高的识别准确率.

本文提出的研究思路和方法的创新之处有以下几点:

1)将 ResNet50 网络模型中的普通 3×3 卷积替换成动态卷积.针对不同身份行人图片的输入,使用不同的卷积核,并对这些不同的卷积核进行注意力加权,从而提升模型准确率.

2)对现有 ResNet50 网络引入自注意力机制与交叉注意力机制,强调行人图片自身特征,忽略一些不必要的错误特征从而提升模型准确率.

3)在 DukeMTMC-ReID、Market-1501 和 MSMT17 数据集上分别进行实验,同时与主流网络模型进行比较.结果表明,本文所提网络模型优于现有模型,具有更卓越的识别性能.

1 相关工作

传统的行人重识别技术主要是对行人的行为进行建模,并在此基础上对其行为进行仿真.段炼等[13]利用贝叶斯模型对时间和空间的位置进行预测,并将时间和空间上的语义信息融合在一起,建立了行人运动特性的数学模型.Helbing 等[14]提出一种“社会力量”模型,即利用“吸引”与“排斥”两种行为模式来描述行人行为.Trautman 等[15]开发了一个互动的 Gauss 进程,这是一个以 Gauss 进程为基础,用来估算群体互动的非参量统计模式.但是,该算法建立在人为设定的特性或一定的规则基础上,在较复杂的情况下需人工调节才能得到较好的效果.同时,该算法的计算复杂性也使其很难应用到大规模、高实时性的应用中.近年来,数据驱动的深度学习算法取得了较好的效果.ResNet 作为残差卷积网络被广泛用于目标分类等领域,在此基础上产生了基于 ResNet 的行人图像检测方法.在行人重识别任务中,由于传统 ResNet 网络只包含静态卷积,卷积形态固定,针对每一个不同的输入图片都只经过同一个卷积,因此特征提取能力较弱.针对这一缺陷,Yang 等[16]提出一种动态滤波器,不同于标准卷积,动态滤波器利用额外的自网络对每个像素生成滤波器,并且采用解耦动态滤波器(Decoupled Dynamic Filter,DDF),在解决自适应的同时比传统卷积更轻量.但是,该方法只有在背景噪声较小的情况下才能识别行人特征信息.冉瑞生等[17]首次证明了当数据足够大时,TransFormer 结构模型可以达到最先进的图片分类精度.但是,与 ResNet 模型相比,基于 TransFormer 的模型往往忽视了行人局部特征,并且缺乏尺度变化、位置编码等信息.因此,本文提出使用动态卷积替换原始网络中的静态卷积,针对不同的输入生成不同的动态卷积核,使网络模型更加灵活高效.

在行人重识别任务中,许多基于注意力的方法被用来提取行人特征.Song 等[18]采用视觉注意机制,将人从背景中分离出来,仅提取人的特征,消除了背景带来的噪声.Franco 等[19]利用卷积注意模块,借助人体姿态信息来定位行人关键部位,提取局部特征向量最终与全局特征向量融合用于分类.尽管此类算法能够在某种程度上缓解由于人体姿势改变而带来的辨识问题,但是多数算法仍需借助人体姿势与骨骼特征点模型,且对模型本身的性能有很大的影响.本文提出将自注意力机制、交叉注意力机制嵌入骨干网络并且同时作用,使模型更聚焦于输入图片本身,减少背景噪声影响,对不同的状态特征给予不同程度的关注,在丰富行人特征的同时,使模型发挥其应有的识别性能,提取行人之间的交互信息,最大可能满足现实需求.

基于距离度量学习的行人重识别方法同样也是目前较为流行的方法之一.其核心思想是,将行人重识别视为聚类问题,以应对相同身份行人图像的挑战.李明哲[20]采用孪生卷积神经网络(Siamese CNN),通过将两个输入图像送入网络,比较它们的特征表示进而学习到两者之间的相似性.在具体实现中,当网络输入是一对身份相同的正样本时,Siamese CNN 的目标是尽可能减小两者特征向量之间的欧氏距离;当输入为一对身份不同的负样本时,网络的目标是尽可能增大这两者特征向量之间的欧氏距离.通过这样的训练方式,网络能够有效地学习到行人图像的特征表示,并在测试时通过比较特征向量来判断图像之间的相似性,从而实现行人重识别的任务.宋婉茹等[21]引入的三重损失是度量学习中被广泛采用的方法,与中心损失相结合使不同种类的数据能够保持一定的距离,从而提高特征的分辨能力.本文采用交叉熵损失函数和难样本三元损失函数共同作用,可以减少行人位移偏差,进而减少因识别而产生的行人特征信息误差和丢失,提高模型识别准确率.

综上,本文提出一种结合动态卷积和注意力机制的行人多特征融合分支网络,并利用交叉熵损失函数和难样本三元损失函数协同作用来降低误差,用于识别行人特征,判断行人身份.

2 Dy-ResNet50 与注意力机制算法

本节主要介绍本文所提出的网络模型,包括网

络模块以及训练模型时用到的损失函数.

2.1 算法概述

本文设计了一个由骨干网络 ResNet50 与 3 个分支组成的行人重识别网络模型,如图 1 所示.本文将 ResNet50 作为原始网络,并将初始网络中的平均池化层(GAP)和全连接层(FC)删除,利用动态卷积替换网络中的普通 3×3 卷积.需要指出的是,本文只将传统 ResNet50 中 Stage1、Stage2 的 Bottleneck 普通卷积替换成动态卷积,并保持 Stage3、Stage4 中的卷积不变,组成具有更高鲁棒性的 Dy-ResNet50 模型.在行为识别过程中,由于图像整体特征和人体局部特征的关注点不同,前者注重整体信息,而后者聚焦于行人的局部关键点,这使得传统的 concatenate( )、average( )、max( )等特征融合方法难以有效整合两者.为了提升网络整体性能,引入注意力机制成为一种合理选择.因此,采用图像全局和人体局部两个自注意力机制,以提取图像特征和人体局部特征的有效信息.通过引入交叉注意力机制,实现了对特征的有效融合,这被认为是一种更切实可行的多特征融合策略.此外,为了避免不合适的池化方法在提取特征时可能导致特征信息丢失,本文在分支 1 中舍弃了传统池化层,选择了一种简单有效的局部重要性池化层(LIP),通过学习一种适应性的权重来增强下采样过程中的判别性信息,从而高效提取行人特征.最后将不同分支得到的特征输入全连接层(FC)进行学习,作为最终的行人身份分类依据.

将 3×160×64 大小的行人特征图输入网络.首先经过骨干网络的第 1 个卷积层(卷积核大小为 7×7)和全局最大池化层(GMP),得到 64×40×16 维特征图,之后经过 Stage1、Stage2 两层动态卷积层,得到 512×20×8 维特征图.本文网络模型中的分支 1 将 Stage2 的输出(512×20×8 维特征图)作为输入.由于低层卷积获得的特征信息在抽取行人图像中的局部特征时相关性不够紧密,于是将人体局部自注意力机制加入骨干网络,利用注意力模块中的键(Key)、查询(Query)和值(Value)3 个向量来计算输入行人图片之间的相关性,再通过局部重要性池化层(LIP)得到 2 048×1×1 维特征向量,由此,分支 1 提取了行人的局部特征.在 Stage4 后面设置独立分支 2.分支 2 将 Stage4 的输出通过图像全局自注意力机制与平均池化层(GAP)的共同作用,得到 2 048×1×1 维特征向量,由此提取行人图片的全局特征.分支 3 是将分支 1 与分支 2 得到的 2 048×1×1 维特征向量作为输入,经过交叉注意力机制、全局最大池化层(GMP),得到结合了行人局部特征与图像全局特征的 2 048×1×1 维特征向量.最后,将不同分支得到的特征向量输入到全连接层(FC)进行融合,批量作用于交叉熵损失函数(Cross Entropy Loss)和难样本三元损失函数(TriHard Loss)进行特征约束.在本文所提网络模型中训练阶段,3 个分支相互监督;测试阶段,将 3 个分支获得的特征向量进行拼接,作为输入图片的特征图,以便后续检验.

2.2 动态卷积

在行人重识别任务中,不同时刻光线强弱变化会导致拍摄照片的明暗不同,行人的肢体形态差异以及不同物体的遮挡都会导致图片中目标特征发生变化.对于传统卷积,卷积核参数对所有输入的行人特征图一视同仁,限制了模型的卷积层数与通道数,

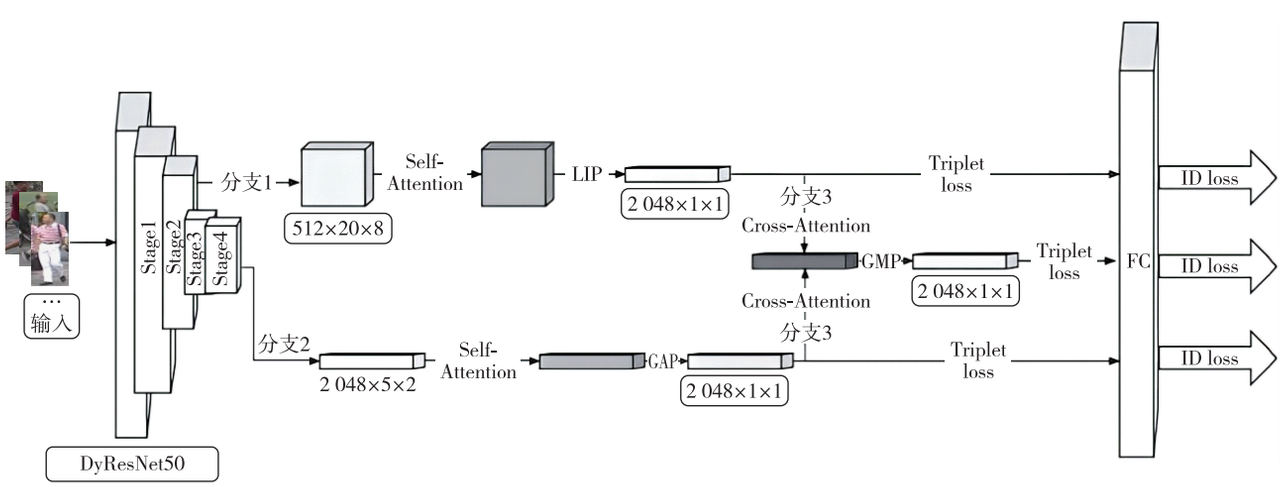

图 1 模型结构

Fig. 1 Model structure

难以满足模型训练所需性能.因此,本文提出使用动态卷积替换 ResNet50 网络模型中的普通 3×3 卷积,以提升模型识别性能,提高模型识别准确率.

对于输入的不同行人特征图,其对应的动态卷积核[22]为

$$O_{\text{output}}(x)=\sigma((\alpha_1 W_1+\cdots+\alpha_n W_n)*x). \quad (1)$$

其中:$\alpha_i=r_i(x)$ 是一个样本依赖加权参数,$*$ 代表卷积.由此,每输入一张不同身份行人的特征图就可计算出与之对应的动态卷积核,在动态卷积中,每层有 K 个卷积核,每个卷积核 W_i 具有与传统卷积核相同的维度.动态卷积利用注意力动态地聚合多个并行卷积核(图 2).在对行人特征图进行处理时,注意力会根据输入行人图片的不同,通过对卷积核进行动态调整,以达到自适应目的.

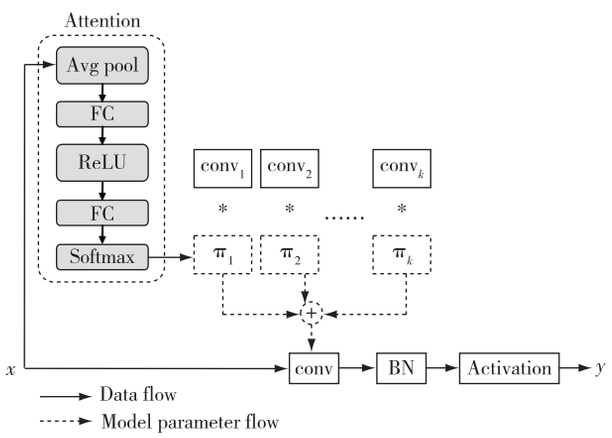

图 2　动态卷积

Fig. 2　Dynamic convolution

本文的动态卷积感知机引入了注意力模型和卷积核的叠加,由平均池化层和两层全卷积层组成,计算量很小,而由于内核体积较小,多个卷积核的运算效率也很高.所以,由动态卷积引起的附加运算量很小,适用于本文所提出的神经网络模型.

传统卷积表达式为 $y=g(W^{\mathrm{T}}+b)$,而本文的动态卷积感知机[23]可表达为

$$y=g(\tilde{W}^{\mathrm{T}}(x)+\tilde{b}(x)), \quad (2)$$

$$\tilde{W}(x)=\sum_{K=1}^{K}\pi_k(x)\tilde{W}_k,\tilde{b}(x)=\sum_{k=1}^{k}\pi_k(x)\tilde{b}_k,$$

$$\text{s.t.}\quad 0\leqslant\pi_k(x)\leqslant 1,\sum_{k=1}^{k}\pi_k(x)=1. \quad (3)$$

其中:$\tilde{W}_k,\tilde{b}_k$ 表示 K 个网络的权重参数;π_k 表示注意力权重.

2.3　图像全局自注意力机制

在实际场景中,同一行人在不同摄像头下可能呈现显著的差异,如图 3 所示.在第一个摄像头中行人背后携带书包,而在第二个摄像头中则看不到.此外,行人的面部特征也可能发生变化,如果模型将该行人的脸部特征以及穿戴信息特征以相同比例跟其他部分特征一起加入身份识别过程,很大程度上会降低模型的识别准确率.为了让计算机也拥有同人眼一样的特征提取能力,本文提出将自注意力机制嵌入原始 ResNet50 网络.

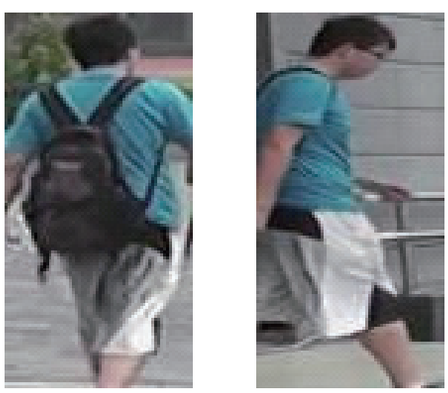

图 3　不同摄像头下同一行人对比

Fig. 3　Comparison of the same person under different cameras

注意力机制的核心思想是对模型输入的各个输入分量赋予不同的权值,使其在特征提取中依据不同的权值给予不同程度的关注.通过对整个图像进行全局自注意力机制的分析,可以确定各个部分对最终分析结果的权值影响.这种方法能够有效消除冗余信息对分析结果的干扰.

在网络模型的第二分支中引入图像全局自注意力机制[24],其结构如图 4 所示.其中,K 代表键(Key),Q 代表查询(Query),V 代表值(Value),它们分别通过可学习的线性映射函数 φ,η,θ 进行特征处理,而 MatMul 则是指矩阵相乘的函数.

将特征映射分别用于生成 Query、Key、Value,这是自注意力机制的关键步骤.通过线性变换,每个位置的特征映射被映射成 3 种表示,用于计算注意力.对于每个位置的 Query 和 Key,计算它们之间的相似度.通常使用点积等方法来计算,然后进行归一化,得到注意力权重,这一步决定了一个位置对其他位置的关注程度:

$$S_v(Q_v,K_v)=\text{softmax}(\frac{Q_v K_v^{\mathrm{T}}}{\sqrt{d_k}}). \quad (4)$$

式中:d_k 表示键向量的维度,决定了控制注意力参数的分布范围.

将计算得到的注意力权重应用到对应位置的

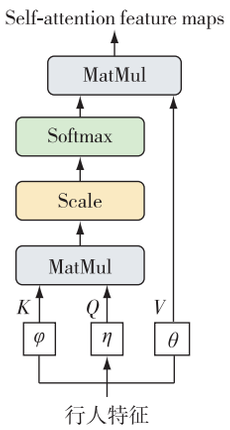

图 4　图像自注意力机制

Fig. 4　Image self-attention mechanism

Value 上,得到加权求和的结果.这表示每个位置对最终表示的贡献,权衡了不同位置的信息:

$$A_{\text{attention}}(Q_v,K_v,V_v)=Z_v=S_v(Q_v,K_v)V_v^{\mathrm{T}}. \tag{5}$$

将加权求和的结果送入后续的网络层进行处理,从而形成最终的图像全局特征:

$$F'_v=F_v+W_vZ_v. \tag{6}$$

式中:W_v 表示图像特征的可学习注意力权值.

2.4　人体局部自注意力机制

通过卷积神经网络等方式对输入的行人图像进行特征提取,得到行人的全局特征表示.本文引入的图注意力机制[25],类似于图像整体自注意力机制,人体局部自注意力机制将行人图像划分为局部区域,例如头部、胸部、下半身等,每个局部区域都被视为一个关键的部分.对每个局部区域的特征进行线性变换,生成关键点特征的键、查询和值被记为 $K_s=Q_s=V_s=\boldsymbol{D}F_s$,其中,$\boldsymbol{D}$ 是可学习的线性转换矩阵,人体关键点特征的可学习注意力权值记为W_s.

对于任意人体关键点特征,可以利用邻近节点特征计算注意力得分,该得分通过权值为 a 的单层前馈神经网络计算获得.

$$S_s=\mathrm{softmax}(\delta(a^{\mathrm{T}}[\boldsymbol{D}f_i\,\|\,\boldsymbol{D}f_j])). \tag{7}$$

其中:符号 ‖ 表示拼接操作;$\delta(\cdot)$ 为激活函数.

图注意力机制处理后得到的关键点特征如下:

$$Z_s=\left\{\delta\left(\sum_{v_j\in B_i}S_s\boldsymbol{D}f_i\right)\right\}. \tag{8}$$

最后,将整合后的表示送入后续的网络层进行处理作为最终的行人表示结果:

$$F'_s=F_s+W_sZ_s. \tag{9}$$

2.5　交叉注意力机制

为了更有效地整合图像信息中的多尺度特征,本文使用交叉注意力机制来精炼和融合两个独立通道的特征,以使两者的信息分布更加合理.为方便说明不同维度的照片信息,以下使用图 5 进行说明.

图 5　输入图片特征分解示意图

Fig. 5　Schematic of input picture feature decomposition

对于同一张行人照片,将人体关键节点特征信息由绿色方框标出,将全局特征信息由黄色方框标出.由图 1 模型结构可知,分支 2 为图像全局特征信息的集合,分支 1 为行人局部特征信息的集合.

将计算得到的注意力权重应用到对应位置或通道上,得到加权求和的结果.这表示每个位置或通道对最终表示的贡献,通过交叉注意力机制融合后的图像特征 Z'_v 和人体关键点特征 Z'_s,可以表示为

$$\begin{cases}Z'_v=S_{v\leftarrow s}(Q_v,K_s)V_v,\\ Z'_s=S_{s\leftarrow v}(Q_s,K_v)V_s.\end{cases} \tag{10}$$

最后,将整合后的表示送入后续的网络层进行处理得到的图像融合特征 F''_v 与人体关键点融合特征 F''_s 为

$$\begin{cases}F''_v=F_v+W_vZ'_v,\\ F''_s=F_s+W_sZ'_s.\end{cases} \tag{11}$$

使用交叉注意力机制可以使得网络能够更灵活地融合不同部分或通道的特征,动态地调整关注度,针对不同维度的行人特征信息进行交叉计算,更好地整合行人特征,从而提高模型对图像或特征的表达能力,在保证较低计算复杂度的同时获得更高的分类准确度.

2.6　损失函数

根据图 1 网络模型的设计以及距离度量的学习,本文采用交叉熵损失函数(Cross Entropy Loss)和难样本三元损失函数(TriHard Loss)共同作为行人重识别任务的损失函数.

对于分支 1 的行人局部特征,网络采用交叉熵损失函数进行训练.在聚合模型中,交叉熵损失函数是最常见的损失函数,它在行人重识别任务中也发挥着重要的作用,其分类数量 N 为训练集中行人身份数,y 为行人真实标签,p_i 为该行人属于第 i 类的预测概率,ϵ 为平滑系数.经过 softmax 后计算交叉熵损失;测试时,舍弃 softmax,使用 3 分支拼接得到的特征向量进行检索.公式表达如下:

$$L_c = -\sum_{i=1}^{N} q_i \log p_i,$$

$$q_i = \begin{cases} \dfrac{\epsilon}{N}, & i \neq y; \\ 1 - \dfrac{N-1}{N}\epsilon, & i = y. \end{cases} \tag{12}$$

对于分支 2 的图像全局特征与分支 3 的融合特征,本文采用难样本三元损失函数进行训练.这是一种基于最优距离测度的损失函数,采用“缩小正向”和“推挤负向”两种方法优化行人特征提取结果.在一个训练(batch)中,随机选择 P 个具有不同身份的行人,每个行人抽取 K 张图片,此时训练样本大小为 $P \times K$.例如对于行人 a,该行人的身份图像为集合 A,剩余图像为集合 B,$d_{a,p}$ 和 $d_{a,n}$ 分别表示正、负样本的距离,α 是函数阈值参数,$(\cdot)_+$ 表示 $\max(\cdot,0)$.公式定义如下:

$$L_{th} = \frac{1}{P \times K} \sum_{s \in \text{batch}} \left(\max_{p \in A} d_{a,p} - \min_{n \in B} d_{a,n} + \alpha \right)_+. \tag{13}$$

测试过程中,将 3 分支获得的行人特征向量进行拼接得到输入图片,利用交叉熵损失函数和难样本三元损失函数评估模型对于新数据的泛化能力,衡量模型识别性能.

因此,模型最终损失函数如下:

$$L = L_c + L_{th}. \tag{14}$$

3 实验分析

3.1 数据集与评价指标

为了验证本文所提网络模型的有效性,本文在 3 个主流公开行人重识别数据集 DukeMTMC-ReID[26]、Market-1501[27] 和 MSMT17[28] 上进行实验评估.

DukeMTMC-ReID 数据集是一个用于行人重识别任务的公共数据集,主要用于评估在多摄像头监控场景下行人重识别算法的性能.该数据集包含来自 8 个不同摄像头的行人图像,涵盖多种日常场景,包括校园、商业区域和户外场景,总共有 1 404 个不同身份的行人,36 411 张图像.数据集被划分为训练集和测试集,训练集包含 702 个不同身份的行人,16 522 张图像;测试集包含 702 个不同身份的行人,包括 2 228 张查询图像和 17 661 张图库图像.

Market-1501 数据集是一个广泛用于行人重识别研究的公共数据集,旨在提供一个丰富而具有挑战性的环境,以评估行人重识别算法在真实场景中的性能.该数据集包含来自 6 个不同摄像头的行人图像,涵盖不同时间和季节的变化,总共有 1 501 个不同身份的行人.每个行人身份都有多张图像,每张图像都被标注了较为详细的信息,包括姿势、视角和背景等.数据集被划分为训练集和测试集,训练集包括 751 个身份,12 936 张图像;测试集包括 750 个身份,19 732 张图像.

MSMT17 数据集是一个大规模、多摄像头的行人重识别数据集,旨在提供更具挑战性和实际场景的数据以促进行人重识别算法的研究.MSMT17 数据集包含来自 15 个不同摄像头的行人图像,总计包含 126 441 张图像.每个行人身份在数据集中都有多张图像,这些图像在姿势、服装和环境等方面都有较大的变化.数据集被划分为训练集和测试集,训练集包括了 32 621 张图像,而测试集包括了 11 659 张查询图像和 82 161 张图库图像,测试集的划分比例为 1∶3.MSMT17 数据集的挑战主要来自于其真实多样的监控场景,包括多摄像头的视角变化、不同天气条件的变化以及行人外观的多样性.该数据集上进行的行人重识别更贴近实际应用.

本文实验将采用平均精度均值(mean Average Precision,mAP)和 Rank-1(第 1 次命中)、Rank-5(第 5 次命中)、Rank-10(第 10 次命中)精度作为模型性能评价指标.

3.2 实验设置

本文实验环境为 64 位 Windows 10 专业版操作系统,算法程序利用 pytorch 1.9.0 深度学习框架,CUDA 11.7、64 GB 内存、24 GB 显存的 NVIDIA GeForce RTX 3090 显卡实现.

在数据处理阶段,对行人图像尺寸统一调整为 160×64.此外,采用图片翻转、对比度增强等操作,以进行数据增强.参数优化选择 Adam 优化器,每个训练批次大小设置为 32,每个测试批次大小设置为 100.为防止数据过拟合,实验共训练 60 个 epoch,初始学习率设置为 0.000 3 以防止学习率过大导致模

型难以收敛,每隔 10 个 epoch 计算一次 mAP、Rank-1、Rank-5 和 Rank-10,在 3 个数据集上保持以上相同的实验设置.

3.3 实验结果与分析

采用公开数据集 DukeMTMC-ReID、Market-1501 和 MSMT17 对本文模型以及主流深度学习模型 DenseNet[29]、SE-ResNet[30]、NasNet[31]、ShuffleNet V2[32]、HACNN[33]、MLFN[34]、OSNet[35] 进行对比实验.表 1 为主流模型与本文模型在 DukeMTMC-ReID 数据集上的对比实验结果.从表 1 中可以看出,本文模型在所有对比模型中取得了最优的分类结果,原因是本文模型中加入的自注意力机制与交叉注意力机制减少了行人特征提取过程中的精度损失.与主流模型相比,在数据集 DukeMTMC-ReID 上本文模型的 Rank-1 与 mAP 较精度最高的 OSNet 算法分别提升 0.9 和 1.6 个百分点.

表 1 不同模型在 DukeMTMC-ReID 上的对比实验

Table 1 Comparative experiment of different models on DukeMTMC-ReID %

模型	Rank-1	Rank-5	Rank-10	mAP
DenseNet	82.0	83.3	84.8	73.1
SE-ResNet	82.3	83.9	85.0	73.4
NasNet	83.2	84.1	85.3	74.1
ShuffleNet V2	84.7	84.3	85.7	74.2
HACNN	80.5			63.8
MLFN	85.3	86.7		77.2
OSNet	87.5	89.7	90.5	77.6
本文模型	**88.4**	**93.1**	**95.2**	**79.2**

表 2 为主流模型与本文模型上在 Market-1501 数据集上的对比实验结果,在数据集 Market-1501 上本文所提模型的 Rank-1 与 mAP 较精度最高的 OSNet 算法分别提升了 0.4 和 0.5 个百分点.

表 2 不同模型在 Market-1501 上的对比实验

Table 2 Comparative experiment of different models on Market-1501 %

模型	Rank-1	Rank-5	Rank-10	mAP
DenseNet	91.0	91.9	93.4	82.9
SE-ResNet	91.2	92.1	94.2	83.1
NasNet	91.4	92.7	94.6	83.9
ShuffleNet V2	92.1	94.3	95.7	84.2
HACNN	91.2			72.6
MLFN	93.2	94.9		81.9
OSNet	94.1	95.3	96.5	84.6
本文模型	**94.5**	**96.0**	**96.8**	**85.1**

表 3 为主流模型与本文模型在 MSMT17 数据集上的对比实验结果,在数据集 MSMT17 上本文所提模型的 Rank-1 与 mAP 较精度最高的 OSNet 算法分别提升了 0.5 和 0.9 个百分点.

表 3 不同模型在 MSMT17 上的对比实验

Table 3 Comparative experiment of different models on MSMT17 %

模型	Rank-1	Rank-5	Rank-10	mAP
DenseNet	75.5	80.3	82.3	45.8
SE-ResNet	77.2	80.9	82.7	52.3
NasNet	77.2	83.5	84.1	53.1
ShuffleNet V2	76.7	83.3	84.7	51.8
HACNN	71.7			42.6
MLFN	76.3	84.7		51.4
OSNet	78.7	85.9	86.6	52.9
本文模型	**79.2**	**86.4**	**88.7**	**53.8**

3.4 消融实验

本文模型采用的骨干网络为传统 ResNet50,为验证引入动态卷积的有效性,选取分支网络的分支 2 参与消融实验.实验结果如表 4 所示,引入动态卷积的 Dy-ResNet50 在分支 2 上的 Rank-1 和 mAP 均优于传统 ResNet50,且选择分支 2 通过图像全局特征进行消融实验,结果更客观,避免了实验的偶然性.因此,选择将动态卷积嵌入传统 ResNet50,组成性能更为优越的 Dy-ResNet50 网络.

表 4 动态卷积在不同分支上的消融实验

Table 4 Ablation experiment of dynamic convolution on different branches %

网络	Rank-1	Rank-5	Rank-10	mAP
ResNet50	76.8	89.5	94.9	54.6
Dy-ResNet50	80.7	92.6	96.4	57.7

与传统卷积不同,动态卷积在每一层都存在 K 个卷积核,模型会利用注意力机制去结合不同卷积核的信息,从而提取到更加丰富的行人特征.实验采用 Market-1501 数据集,令 K 为 2,4,6,8,结果如图 6 所示.从图 6 中可以看出:当卷积核的个数太少时,会导致模型的特征抽取不够充分,降低模型的识别精度;当卷积核数目过多时,会导致网络模型趋于复杂,识别精度得不到提升.在动态卷积核数 K 为 4 的情况下,所得到的网络模型表现最佳.

为验证动态卷积核在 ResNet50 不同层的效果,将动态卷积分别添加到 Stage1、Stage2、Stage1+Stage2

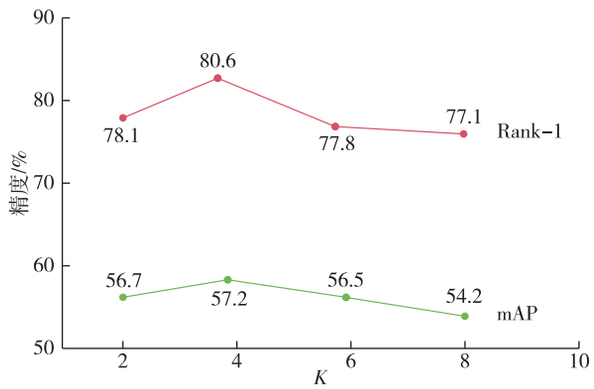

图 6　不同动态卷积核数目对模型性能的影响

Fig. 6　Influence of number of dynamic convolution kernels on model performance

中,实验采用 Market-1501 数据集,结果如表 5 所示.从表 5 可知,将动态卷积联合作用于 Stage1+Stage2 的 Rank-1、Rank-5、Rank-10 和 mAP 都要优于单独作用于其中某一层,同时由于本文引入的动态卷积内核较小,因此,作用于 Stage1+Stage2 模型训练时间无显著增加,模型训练效率几乎不受影响,所以,在该模型中使用动态卷积核是可行的.

表 5　ResNet50 在不同层利用动态卷积的消融实验

Table 5　Ablation experiment using dynamic convolution in different layers of ResNet50　%

方法	Rank-1	Rank-5	Rank-10	mAP
Original	77.5	90.2	95.1	56.4
Stage1	79.1	91.6	95.3	56.9
Stage2	79.6	91.8	95.6	57.0
Stage1+Stage2	81.1	92.5	97.1	57.8

本文通过对公共数据集 DukeMTMC-ReID、Market-1501 和 MSMT17 进行消融实验,以验证不同注意力模块的有效性,实验结果如表 6、7、8 所示.可见,对于图像全局特征和人体局部特征,引入自注意力模块都能提高模型的识别准确率,证明了自注意力机制的有效性.鉴于图像全局特性和人体局部特征之间存在显著差异,直接拼接这两类特征并嵌入单一自注意力机制并不具备明显的优势.虽然引入自注意机制后可以有效减少冗余信息,提高识别精度,但与 OSNet 模型相比,识别精度仍有改进空间.通过引入交叉注意力机制,成功实现了更有效的图像全局特征和人体局部特征融合,从而显著提升了识别准确率.

表 6　在 DukeMTMC-ReID 数据集上注意力机制消融实验结果

Table 6　Experimental results of attention mechanism ablation on DukeMTMC-ReID　%

Dy-ResNet50	图像自注意力机制	人体局部自注意力机制	交叉注意力机制	Rank-1	mAP
√	×	×	×	82.3	73.4
√	√	×	×	85.9	77.4
√	√	√	×	86.7	78.0
√	×	√	×	86.5	77.9
√	√	√	√	88.4	79.2

表 7　在 Market-1501 数据集上注意力机制消融实验结果

Table 7　Experimental results of attention mechanism ablation on Market-1501　%

Dy-ResNet50	图像自注意力机制	人体局部自注意力机制	交叉注意力机制	Rank-1	mAP
√	×	×	×	90.3	82.1
√	√	×	×	93.4	83.2
√	√	√	×	93.1	83.0
√	×	√	×	93.6	83.7
√	√	√	√	94.5	85.1

表 8　在 MSMT17 数据集上注意力机制消融实验结果

Table 8　Experimental results of attention mechanism ablation on MSMT17　%

Dy-ResNet50	图像自注意力机制	人体局部自注意力机制	交叉注意力机制	Rank-1	mAP
√	×	×	×	75.7	46.2
√	√	×	×	76.2	51.3
√	√	√	×	76.8	52.0
√	×	√	×	76.0	51.1
√	√	√	√	79.2	53.8

为了能更直观地展现本文模型识别行人身份效果,图 7 展示了 Market-1501 数据集对应 Rank-1 到 Rank-10 的查询结果,其中,黑色框对应的是查询图像,绿色框对应的是正确查询结果,红色框对应的是错误查询结果.由图 7 所示,依托于自注意力机制与交叉注意力机制联合作用的网络模型在前 5 个查询结果中大致可以正确地识别出 4 个行人身份,证明了本文模型具有较高的识别准确率和身份识别能力.

4　结束语

本文提出了一种基于动态卷积与注意力机制的多特征融合分支网络模型.该模型主要由骨干网络

图 7　可视化结果

Fig. 7　Visualized results

ResNet50 与 3 个分支组成,将 ResNet50 中前两个 Bottleneck 的 3×3 卷积替换成动态卷积,Stage2、Stage4 的输出分别作为分支 1、分支 2 的输入,同时在分支 1 与分支 2 中引入自注意力机制进行不同分支的行人图片特征提取,并引入交叉注意力机制作为分支 3,最终将各分支提取的行人有效特征进行融合.实验结果表明,本网络模型在公共数据集 DukeMTMC-ReID、Market-1501 和 MSMT17 上均取得了不错的识别效果.未来将进一步扩展数据集,研究利用 3D 模型提取行人特征,更好地应用到现实网络中.

参考文献

References

[1] 郭彤,赵倩,赵琰,等.多分支融合注意力机制的行人重识别方法[J].计算机工程与设计,2022,43(8):2260-2267
GUO Tong, ZHAO Qian, ZHAO Yan, et al. Person re-identification method based on multi-branch fusion attention mechanism[J]. Computer Engineering and Design, 2022,43(8):2260-2267

[2] Sun Y F, Zheng L, Yang Y, et al. Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline)[M]//Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 501-518

[3] Fu Y, Wei Y C, Zhou Y, et al. Horizontal pyramid matching for person re-identification[J]. arXiv e-Print, 2018, arXiv:1804.05275

[4] Wang F, Mao R S, Yan L F, et al. A deep learning-based approach for rectus abdominis segmentation and distance measurement in ultrasonography[J]. Frontiers in Physiology, 2023, 14:1246994

[5] Sun M, Wang Y F, Zeng M Q, et al. Development and application of creepage distance measurement system for zinc oxide arrester[J]. Journal of Physics: Conference Series, 2023, 2591(1):012046

[6] 张聪聪,何宁.基于关键帧的双流卷积网络的人体动作识别方法[J].南京信息工程大学学报(自然科学版),2019,11(6):716-721
ZHANG Congcong, HE Ning. Human motion recognition based on key frame two-stream convolutional network [J]. Journal of Nanjing University of Information Science & Technology (Natural Science Edition), 2019, 11(6): 716-721

[7] 李金轩,杜军平,周南.基于注意力特征提取网络的图像描述生成算法[J].南京信息工程大学学报(自然科学版),2019,11(3):295-301
LI Jinxuan,DU Junping,ZHOU Nan.Image caption algorithm based on an attention image feature extraction network[J].Journal of Nanjing University of Information Science & Technology (Natural Science Edition),2019,11(3):295-301

[8] 刘忠洋,周杰,陆加新,等.基于注意力机制的多尺度特征融合图像去雨方法[J].南京信息工程大学学报(自然科学版),2023,15(5):505-513
LIU Zhongyang,ZHOU Jie,LU Jiaxin,et al.Image rain removal via multi-scale feature fusion based on attention mechanism[J].Journal of Nanjing University of Information Science & Technology (Natural Science Edition),2023,15(5):505-513

[9] Wang J Y,Jang J S R.Training a singing transcription model using connectionist temporal classification loss and cross-entropy loss[J].IEEE/ACM Transactions on Audio,Speech,and Language Processing,2022,31:383-396

[10] Yang Z,Yuan Y,Xu Y,et al.FACE:evaluating natural language generation with Fourier analysis of cross-entropy[J].arXiv e-Print,2023,arXiv:2305.10307

[11] Cheng D,Gong Y H,Zhou S P,et al.Person re-identification by multi-channel parts-based CNN with improved triplet loss function[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).June 27-30,2016,Las Vegas,NV,USA.IEEE,2016:1335-1344

[12] Bui T,Ribeiro L,Ponti M,et al.Compact descriptors for sketch-based image retrieval using a triplet loss convolutional neural network[J].Computer Vision and Image Understanding,2017,164:27-37

[13] 段炼,胡涛,朱欣焰,等.顾及时空语义的疑犯位置时空预测[J].武汉大学学报(信息科学版),2019,44(5):765-770
DUAN Lian,HU Tao,ZHU Xinyan,et al.Spatio-temporal prediction of suspect location by spatio-temporal semantics[J].Geomatics and Information Science of Wuhan University,2019,44(5):765-770

[14] Helbing D,Molnár P.Social force model for pedestrian dynamics[J].Physical Review E,1995,51(5):4282-4286

[15] Trautman P,Krause A.Unfreezing the robot:navigation in dense,interacting crowds[C]//2010 IEEE/RSJ International Conference on Intelligent Robots and Systems.October 18-22,2010,Taipei,China.IEEE,2010:797-803

[16] Yang J R,Zheng W S,Yang Q Z,et al.Video-based temporary volume network re-certification[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR).June 14-19,2020,Seattle,WA,USA.IEEE,2020:3286-3296

[17] 冉瑞生,石凯,江小鹏,等.基于双注意力 CrossViT 的微表情识别方法[J].南京信息工程大学学报(自然科学版),2023,15(5):541-550
RAN Ruisheng,SHI Kai,JIANG Xiaopeng,et al.Micro-expression recognition based on dual attention CrossViT[J].Journal of Nanjing University of Information Science & Technology (Natural Science Edition),2023,15(5):541-550

[18] Song C F,Huang Y,Ouyang W L,et al.Mask-guided contrastive attention model for person re-identification[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.June 18-23,2018,Salt Lake City,UT,USA.IEEE,2018:1179-1188

[19] Franco A,Oliveira L.A coarse-to-fine deep learning for person re-identification[C]//2016 IEEE Winter Conference on Applications of Computer Vision (WACV).March 7-10,2016,Lake Placid,NY,USA.IEEE,2016:1-7

[20] 李明哲.基于时空注意力机制的视频行人再识别方法研究[D].哈尔滨:哈尔滨工程大学,2020
LI Mingzhe.Research on video pedestrian recognition method based on spatio-temporal attention mechanism[D].Harbin:Harbin Engineering University,2020

[21] 宋婉茹,赵晴晴,陈昌红,等.行人重识别研究综述[J].智能系统学报,2017,12(6):770-780
SONG Wanru,ZHAO Qingqing,CHEN Changhong,et al.Survey on pedestrian re-identification research[J].CAAI Transactions on Intelligent Systems,2017,12(6):770-780

[22] 耿韶松,李晋国.基于动态卷积与注意力的多特征融合行人重识别[J].计算机工程与设计,2023,44(4):1228-1234
GENG Shaosong,LI Jinguo.Person re-identification based on multi-feature fusion of dynamic convolution and attention[J].Computer Engineering and Design,2023,44(4):1228-1234

[23] Cheng X,Zhou J M,Zhao X M,et al.A presentation attack detection network based on dynamic convolution and multi-level feature fusion with security and reliability[J].Future Generation Computer Systems,2023,146:114-121

[24] 赵小虎,尹良飞,赵成龙.基于全局-局部特征和自适应注意力机制的图像语义描述算法[J].浙江大学学报(工学版),2020,54(1):126-134
ZHAO Xiaohu,YIN Liangfei,ZHAO Chenglong.Image captioning based on global-local feature and adaptive-attention[J].Journal of Zhejiang University (Engineering Science),2020,54(1):126-134

[25] 饶天荣,潘涛,徐会军.基于交叉注意力机制的煤矿井下不安全行为识别[J].工矿自动化,2022,48(10):48-54
RAO Tianrong,PAN Tao,XU Huijun.Unsafe action recognition in underground coal mine based on cross-attention mechanism[J].Journal of Mine Automation,2022,48(10):48-54

[26] Ristani E,Solera F,Zou R,et al.Performance measures and a data set for multi-target,multi-camera tracking[M]//Lecture Notes in Computer Science.Cham:Springer International Publishing,2016:17-35

[27] Zheng L,Zhang H H,Sun S Y,et al.Person re-identification in the wild[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).July 21-26,2017,Honolulu,HI,USA.IEEE,2017:3346-3355

[28] Wei L H,Zhang S L,Gao W,et al.Person transfer GAN to bridge domain gap for person re-identification[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.June 18-23,2018,Salt Lake City,UT,USA.IEEE,2018:79-88
[29] Huang G,Liu Z,Van Der Maaten L,et al.Densely connected convolutional networks [C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).July 21-26,2017,Honolulu,HI,USA.IEEE,2017:2261-2269
[30] Cai L Q,Li H,Dong W,et al.Micro-expression recognition using 3D DenseNet fused squeeze-and-excitation networks [J].Applied Soft Computing,2022,119:108594
[31] Zoph B,Vasudevan V,Shlens J,et al.Learning transferable architectures for scalable image recognition [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.June 18-23,2018,Salt Lake City,UT,USA.IEEE,2018:8697-8710
[32] Ma N N,Zhang X Y,Zheng H T,et al.ShuffleNet v2:practical guidelines for efficient CNN architecture design [M]//Computer Vision-ECCV 2018.Cham:Springer International Publishing,2018:122-138
[33] Li W,Zhu X T,Gong S G.Harmonious attention network for person re-identification[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.June 18-23,2018,Salt Lake City,UT,USA.IEEE,2018:2285-2294
[34] Chang X B,Hospedales T M,Xiang T.Multi-level factorisation net for person re-identification[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition.June 18-23,2018,Salt Lake City,UT,USA.IEEE,2018:2109-2118
[35] Zhou K,Yang Y,Cavallaro A,et al.Learning generalisable omni-scale representations for person re-identification[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2021,44(9):5056-5069

Multi-feature person re-identification based on cross-attention mechanism

WU Xinyi[1] DENG Zhiliang[1] LIU Yunping[1] DONG Juan[2] LI Jiaqi[1]
1 School of Automation,Nanjing University of Information Science & Technology,Nanjing 210044,China
2 School of Electronics & Information Engineering,Nanjing University of Information Science & Technology,Nanjing 210044,China

Abstract Existing person re-identification (Re-ID) methods often struggle with inaccurate feature extraction and misidentification of person features due to environmental noise.Here,we propose a multi-feature fusion branch network for person Re-ID based on dynamic convolution and attention mechanism.First,considering the uncertainties in illumination,human posture and occlusion,dynamic convolution is proposed to replace static convolution in ResNet50 to obtain a more robust Dy-ResNet50 model.Second,given the great difference in camera perspective and the likelihood of people being occluded by objects,self-attention and cross-attention mechanisms are embedded into the backbone network.Finally,the cross entropy loss function and the hard triplet loss function are used as the model's loss functions,and experiments are carried out on public datasets of DukeMTMC-ReID,Market-1501 and MSMT17.The results show that the proposed model outperforms current mainstream models in Rank-1 (first hit) and mAP (mean Average Precision) on three public datasets,indicating its high identification accuracy.

Key words person re-identification; dynamic convolution; self-attention mechanism; cross-attention mechanism