面向群体行为识别的注意力池化机制

李定^{1,2}, 张文生^{1*}

1. 中国科学院自动化研究所, 北京 100091

2. 中国科学院大学人工智能学院, 北京 100049

* 通信作者. E-mail: zhangwenshengia@hotmail.com

收稿日期: 2020-08-02; 修回日期: 2020-09-26; 接受日期: 2020-11-06; 网络出版日期: 2021-02-25

科技创新 2030 — “新一代人工智能” 重大项目 (批准号: 2018AAA0102100) 资助

摘要 视频行为识别近年来逐渐成为计算机视觉领域学者的研究热点, 按照识别对象进行划分, 视频行为识别任务可分为个体行为识别与群体行为识别. 本文聚焦于群体行为识别, 识别与分析视频场景中整体人群的行为. 已有的群体行为识别方法大多采用多层时序网络模型, 学习得到表征时序变化的个体行为特征并对其进行聚合形成群体行为特征. 但是, 在个体特征聚合过程中, 以往方法未能有效考虑个体对群体行为贡献程度的差异性, 影响识别性能. 为此, 本文提出一种针对个体行为特征聚合的注意力池化机制, 并依此建立了新型群体行为识别模型, 以自底向上的方式同时实现个体行为与群体行为分层识别. 首先利用卷积神经网络提取视频中人体图像区块的个体静态特征, 并将其作为多层递归神经网络时序模型的输入, 从而得到个体动态特征. 随后通过注意力池化机制对个体特征完成聚合, 得到相应的群体行为特征; 最后依托个体、群体行为特征同时完成个体行为与群体行为的识别. 未验证所提方法的有效性, 本文依托广泛使用的 The Volleyball Dataset 数据集上开展了一系列实验验证. 结果显示, 本文所提出的模型取得了较好的分类准确率, 分类性能优于当前先进模型.

关键词 群体行为识别, 表示学习, 注意力机制, 深度学习

1 引言

针对视频行为识别任务, 计算机视觉领域内研究人员提出了许多有效的行为识别方法^[1~6], 但大部分方法仅局限于针对单个个体的行为识别. 而多人情景下发生的事件不仅与个体行为有关, 也与人群整体的行为有很大的联系. 群体行为与单个个体的行为关系密切, 同时也在行为定义、表现等方面具有较大的差异. 对群体行为进行识别不但具有重要的科学意义, 而且在智能安防、智慧城市, 以及视频检索等领域具有重要的应用价值. 近年来, 群体行为识别受到了研究者的广泛关注^[7~13], 其主要难点在于如何对个体与群体的层次关系进行建模.

引用格式: 李定, 张文生. 面向群体行为识别的注意力池化机制. 中国科学: 信息科学, 2021, 51: 399–412, doi: 10.1360/SSI-2020-0235
Li D, Zhang W S. Attentive pooling for group activity recognition (in Chinese). Sci Sin Inform, 2021, 51: 399–412, doi: 10.1360/SSI-2020-0235

图 1 (网络版彩图) 群体行为左方回合获胜示例图
Figure 1 (Color online) Example of left winpoint

图 2 (网络版彩图) 群体行为右方扣球进攻示例图
Figure 2 (Color online) Example of right spike

为解决这一难点, 研究者采用个体级与群体级网络相结合的多层级结构分别对相关行为进行建模, 开展了一系列研究工作 [8, 13, 14]. 群体行为不仅是对多个个体行为的简单聚合, 更重要的是在聚合过程中将个体相关关系紧密结合进来. 因此, 这一系列方法首先在个体层级逐个地提取对个体行为及其变化的特征. 之后, 将个体行为及其时变特征进行聚合得到相应的群体行为特征. 具体来看, LSTM (long-short-term memory recurrent neural network) [15] 等递归神经网络 (RNN) 被用来刻画人体行为及其变化, 而在聚合过程中一般采用常见的最大池化或平均池化的方式. 该类池化方式主要依托群体内成员的行为一致性来进行特征聚合, 例如如图 1 中类别为“左方本回合获胜”的视频片段, 依据该一致性原则, 可判断球网左方的队伍正集中在一起庆祝本回合获胜, 而其对手在本回合落败.

尽管上述方法 [8, 9, 13, 14] 在一定程度上提升了识别性能, 但均在特征聚合过程中使用了最大池化机制, 这使得此类模型未能有效地考虑不同个体对于群体行为贡献的差异性. 在群体行为定义过程中, 往往一部分个体行为会对于群体行为具有较大影响. 例如, 图 2 的视频片段展示的是群体行为“右方扣球进攻”, 注意到球网右方有一名球员的个体行为被标注为“扣球” (见黄色矩形框), 而球网左边的两名球员为进行防守, 主动上前“拦网” (见红色矩形框), 与此同时视野内的其他球员个体行为有的被标注为“移动”, 有的被标注为“站立”或者“等待”, 实时准备下一回合比赛. 观察可得, 我们依靠“扣球”和“拦网”的球员就可对群体行为“右方扣球进攻”进行识别, 这些较为重要的个体对群体行为的

贡献较大,而其他球员贡献相对较小,两者之间具有显著的差异性.个体间差异增大了个体间上下文关系的复杂度,使得最大池化机制无法有效地满足特征聚合的要求.

为此,本文提出了一种注意力池化机制,不仅依托群体内部的一致性来进行简单上下文关系时的特征聚合,而且同时利用个体间差异性实现重要个体的显著性表示和非重要个体的噪声滤除.本文的主要贡献如下:

- 本文提出了一种面向群体行为识别的注意力池化机制,利用群体中的个体差异性与内部一致性实现复杂个体间上下文关系时的特征聚合.
- 本文引入群体层级关系进一步拓展了全局注意力池化机制,提出了多层注意力池化机制,利用衍生的多种池化模式聚合个体行为特征与子群体行为特征.
- 本文依托 The Volleyball Dataset 数据集展开一系列实验验证,实验结果表明所提方法可有效提升群体行为识别性能.

本文的章节介绍如下:第2节围绕群体行为识别与注意力机制,对相关工作进行简要总结回顾;第3节正式提出面向群体行为的注意力池化机制及其变式,并分析了注意力池化模块的内部结构;第4节在通用数据集上进行了消融实验并与当前先进模型进行比较,实验证明了融合注意力池化机制的模型具备最优的性能;第5节介绍了全文结论,并针对未来研究方向提出计划.

2 相关工作

本节聚焦于群体行为识别任务研究现状,并简要介绍深度学习中的注意力机制,文献[16,17]回顾了近年来本领域的代表性工作.

2.1 群体行为识别

群体行为识别任务的核心在于通过表示学习方法得到具有高判别性的群体行为特征.在早期研究工作中,研究者广泛利用先验知识手工设计相应的行为特征,用以对个体与群体行为进行表示. Nahi 等^[18]设计了一种 pose-let activation pattern over time (TPOS) 描述子用以捕捉群体内的个体运动与交互信息. Choi 等^[7]提出一种集成框架,可同时对多人进行跟踪以及分层级行为识别.但以手工设计特征为基础的方法具有较大的劣势,高度依赖研究者对该任务的先验假设,而且无法完成端到端训练.

近年来,随着深度学习的不断发展,深度神经网络被广泛应用于群体行为识别任务中. Ibrahim 等^[8]提出了一种层次化级联的时序神经网络模型,深刻启发了本领域的研究者.基于该类型时序神经网络的方法总体上可划分为3步:(1)首先将视野中的人体行为图像区块进行导出,依据人体轨迹构建行为图像序列,随后将该行为图像序列输入卷积神经网络(CNN)和低层RNN中,提取得到行为对应的空间特征和时序变化特征;(2)在提取得到个体时空特征的基础上,对所有个体特征进行池化聚合,并把聚合结果输入高层RNN中,利用高层RNN中提取群体行为特征;(3)将高层RNN输出的群体行为特征送入全连接层,经过softmax函数得到不同的类别得分.

随后,许多研究者针对这一模型进行了改进与拓展.群体本身可进一步精细划分, Ibrahim 等^[14]将子群体作为研究对象,随后对子群体分类结果进行融合,进一步提高了识别精度. Shu 等^[9]从能量的角度对网络进行了优化,提出了基于能量最小化的新型损失函数,进一步提升了模型性能. Bagautdinov 等^[11]借鉴多任务学习的思想,提出一种多任务框架,在不同的分支进行人体目标检测与群体行为识别. Li 等^[10]深度挖掘了个体行为标注中的语义信息,基于标注语义同时实现个体行为描述与群体行为识别. 结构递归神经网络(SRNN)对于处理具有“群体-个体”层级架构的数据具有较大优势, Biswas

等^[13] 将其引入群体行为识别任务中, 针对视野中的个体行为、交互行为、群体行为分别进行建模. 上述基于深度学习的方法虽然利用模型学习得到了较好的行为表示, 但在个体特征聚合时仍采用最大池化, 忽略了不同个体对群体行为贡献的差异性.

2.2 注意力机制 (attention mechanism)

深度学习中的注意力机制源于人脑感知机理, 重点关注输入中的关键内容, 而忽略低价值的信息, 在计算机视觉和自然语言处理等领域的多个任务中得到了广泛的应用^[19]. 从注意力的处理流程出发, 可分为硬性注意力机制与软性注意力机制.

在硬性注意力机制中, Mnih 等^[20] 在递归神经网络中有效嵌入融合注意力机制, 可从图像中通过自适应的方式选择一系列图像区块来实现区域关注. 但硬性注意力机制的训练依赖强化学习流程, 增加了模型收敛难度. 软注意力机制主要通过使用加权平均的方式来进行选择, 而非对其进行硬性选择. 在软注意力机制中, Sharma 等^[21] 提出了一种建立在多层 RNN 网络基础之上的 Soft-Attention 模型. 此方法通过多次筛查来选择性地关注重要的视频帧, 并完成视频分类. Yang 等^[22] 提出用于文档分类的分层关注网络, 在这种方法中, 网络具有两个级别的注意力机制分别应用于单词和句子级别, 使其能够差异化地关注到相对重要的内容. Li 等^[10] 提出了一种用于视频中动作识别的端到端序列学习模型, 利用带有软注意力机制的 LSTM 来捕捉空间布局. Long 等^[23] 在局部特征聚合过程中采用注意力机制, 并将其引入注意集群视频分类任务中. 此外, Bengio 等^[24] 首先提出了自注意力机制, 受到学术界的广泛关注, 其核心思想在于利用自身特征来有效地实现软性注意力机制. 本文中的注意力池化模块核心思想与自注意力机制类似, 以个体行为特征为输入, 利用个体贡献差异生成相应的注意力权重, 相关参数随模型训练进行更新.

注意力机制使得模型关注于重要区域, 减少噪声负面影响, 从而提高识别性能. 特别需要指出的是, 注意力机制通过为群体中的元素分配相关性分数实现元素重要性评估, 突出显示与任务最相关的元素信息. 此外, 它一定程度上反映了特征集合内部的结构信息, 有利于提升模型可解释性.

3 基于注意力池化的群体行为识别模型

本文以视频中的整体人群行为作为研究对象, 将个体行为与群体行为分层进行表示, 采用一种自下而上的方法同时实现两者的分类识别. 模型针对个体与群体进行了分级处理, 在从个体级特征到群体级特征的聚合过程中采用注意力池化聚合机制. 对群体行为识别的第 1 步就是在个体层级获取个体行为及其时序动态变化的特征, 随后将其输入特征聚合模块得到群体行为特征. 合理利用个体与群体的相互关系可有效增强模型表征能力, 注意力池化的特征聚合模块同时考虑视野中个体间差异性和群体内部的一致性. 模型整体框架如图 3 所示.

以当前视野中的第 i 个人为例, 首先将对应图像区块输入 CNN 中, 提取得到其空间静态特征 $x_i = (x_i^1, x_i^2, \dots, x_i^T)$. 紧接着, 将上一步提取到的空间静态特征 x_i 送入个体级 LSTM 网络中, 将个体级 LSTM 网络隐层输出作为对应的时序特征 $h_i = (h_i^1, h_i^2, \dots, h_i^T)$, 其中, T 代表图像序列长度. LSTM 网络中的参数与整体模型一起完成初始化, 之后随着网络训练进行更新, 时序特征 h_i 的提取过程可简要表示为

$$h_i^t = \text{LSTM}(h_i^{t-1}, x_i^t). \quad (1)$$

为了完整地表示场景内的个体行为, 我们对提取到的空间特征与时序特征在通道维度进行拼接 (表

图 3 (网络版彩图) 基于注意力池化的群体行为识别模型示意图

Figure 3 (Color online) Framework of attentive pooling based model for group activity recognition

示为 $\oplus$). 在 t 时刻, 记个人行为特征为 P_i , 则有

$$P_i^t = h_i^t \oplus x_i^t. \quad (2)$$

在提取得到个体行为特征的基础上, 需在行为特征层面构建起对应的层级关系. 因此, 本文选择对上一步提取得到的个体行为特征进行特征聚合, 由此实现行为信息的层级提升, 得到用来表示群体行为的特征. 为此, 本文提出注意力池化机制, 将个体间贡献差异性引入层级提升的过程中. 紧接着将池化后的特征 G_t 输入群体级的 LSTM 网络, 最后利用该网络隐层状态 H_g^t 作为群体行为特征完成群体行为分类. 在整体网络训练过程中, 我们构建了如下损失函数 (loss function):

$$L = L(\phi_g(H_g^t, y_g^t, \theta_g)) + \lambda \frac{1}{n} \sum_{i=1}^n L(\phi_p(P_i^t, y_i^t, \theta_i)), \quad (3)$$

其中, L 代表交叉熵损失函数, ϕ_g 代表群体行为识别模型, ϕ_p 为个体行为识别模型, y_g^t 和 y_i^t 分别表示群体行为和个体行为的类别标签, n 指当前视野中群体内部所包含的个体数, λ 为个体与群体行为识别训练损失的权重系数.

4 注意力池化机制

在具体场景中, 群体行为由个体行为共同定义得到. 在该条件下, 某些个体行为和对应的群体行为之间存在很强的相关性, 而其余的个体行为对于群体行为仅有较小的影响程度. 注意力池化机制旨在将个体级别的行为信息全面、有区分地进行汇集, 以从而获得更有判别性的群体行为表示, 同时也可以自然地将我们的关注对象从单个个体转换至整个群体.

注意力机制因其关注重点内容的特性, 在计算机视觉和自然语言处理等领域的多个任务中得到了广泛应用, 为实现有效特征聚合提供了新的思路. 本文针对群体行为识别任务设计了集约型注意力池化与拼接型注意力池化两种机制, 在个体行为特征向群体行为特征的过渡阶段利用注意力池化机制进行特征聚合. 前者主要将群体看作一个整体, 最终获取其群体行为特征; 后者将群体拆分为若干子群体, 不同子群体行为特征进行拼接, 从而得到对应的群体行为特征.

4.1 集约型注意力池化

集约型注意力池化的核心思想是对个体层网络输出的多个个体行为特征进行聚合, 最终得到与单个个体特征维数相同的特征, 并输入群体层网络.

4.1.1 全局注意力池化

CNN 在视觉特征提取方面展现出了强大的能力, 因此本文选用 CNN 作为特征提取器, 在每个采样时刻根据当前状态下人体所处位置信息提取其周围图像区块的空间特征. 与文献 [8] 相似, 本文使用 LSTM 网络来提取个人行为的时序动态变化特征. 最终, 个人行为的特征由时序动态信息和空间静态特征共同构成. LSTM 被广泛用于多个领域中的序列问题, 其内部含有多个控制信息流量的门控单元, 可以根据序列变化对信息流向进行长短时记忆控制. 当输入序列较长时, 含信息流控制的神经网络可较好地避免训练 RNN 过程中的梯度弥散, 也便于对其进行分层堆叠.

与前文类似, 对于 t 时刻下视野中的第 i 个人来说, 其个体行为特征为 P_i^t , 随后将 P_i^t 输入至含一层隐层网络的多层感知机 (MLP) 中, 得到输出 u_i^t ; 接着计算 u_i^t 与上下文向量 u_p^t 之间的相似度. 其中, u_i^t 为 P_i^t 的隐层表示, u_p^t 表示个体级别的上下文向量. 通过计算 u_i^t 与 u_p^t 之间的相似性来对个体的重要性程度进行量化, 从而将个体间差异性引入特征聚合过程中, 使加权平均后的个体行为特征更好地反映群体行为的特性. 全局注意力池化 (global attentive pooling, GAP) 计算过程如下:

$$u_i^t = \tanh(W_p^t P_i^t + b_p^t), \quad (4)$$

$$\alpha_i^t = \frac{\exp(u_i^t \cdot u_p^t)}{\sum_i \exp(u_i^t \cdot u_p^t)}, \quad (5)$$

$$G_t = \sum_{i=1}^n \alpha_i^t P_i^t, \quad (6)$$

其中, α_i^t 是对应于个体特征 P_i^t 的权重, 假设视野中的人群中包含 n 个个体, 全局注意力池化结构如图 4 所示.

4.1.2 多层注意力池化

群体内部本身存在客观的组织结构信息, 个体与群体之间存在客观从属关系, 组织结构关系的引入有利于细粒度地进行特征聚合. 因此, 本文将整个群体划分为若干个子群体, 每个子群体内部包含若干个体. 在引入子群体层级的情况下, 整个群体的行为由个体行为与子群体行为共同定义而成. 引入子群体之后, 不同子群体的行为对整个群体行为的影响同样具有差异性. 例如, 对于群体行为“右方队伍本回合获胜”, 视野中球网右侧的子群体的重要性通常显著高于球网左侧的子群体. 因此, 本文在全局注意力池化机制的基础上进一步拓展, 设计了多层注意力池化机制, 对子群体行为特征及其内部的个体特征均进行注意力池化聚合.

首先, 我们将视野中的人群划分为多个子群体. 随后, 分别在各子群体内部对个体行为特征使用全局注意力池化机制进行聚合, 由此生成子群体级别的行为特征. 之后, 对上一步得到的各子群体级行为

图 4 (网络版彩图) 全局注意力池化示意图

Figure 4 (Color online) Global attentive pooling at time-step t

特征进行注意力池化, 输出为 G_t . 对于第 j 个子群体中的第 i 个个体来说, 个体级注意力池化 (person attention) 计算过程可表示为

$$u_{ij}^t = \tanh(W_{pj}^t P_{ij}^t + b_{pj}^t), \quad (7)$$

$$\alpha_{ij}^t = \frac{\exp(u_{ij}^t \cdot u_{pj}^t)}{\sum_i \exp(u_{ij}^t \cdot u_{pj}^t)}, \quad (8)$$

$$g_j^t = \sum_{i=1}^{n_j} \alpha_{ij}^t P_{ij}^t. \quad (9)$$

对于第 j 个子群体来说, 子群体级的注意力池化 (subgroup attention) 计算过程可表示为

$$u_j^t = \tanh(W_g^t g_j^t + b_j^t), \quad (10)$$

$$\alpha_j^t = \frac{\exp(u_j^t \cdot u_g^t)}{\sum_j \exp(u_j^t \cdot u_g^t)}, \quad (11)$$

$$G_t = \sum_{j=1}^m \alpha_j^t g_j^t, \quad (12)$$

其中, u_j^t 为 g_j^t 的隐层表示, u_g^t 为子群体级别的上下文向量. 多层注意力池化 (hierarchical attentive pooling, HAP) 整体结构如图 5 所示.

4.2 拼接型注意力池化

拼接型注意力池化 (attentive pooling with subgroups concatenation, APSC) 是根据“群体-子群体-个体”组织结构, 将多个个体行为特征进一步聚合、缩减, 最终得到较少的多个特征并对多个特征进行拼接.

本文将拼接后的各子群体行为特征输入至群体级别的 LSTM 时序网络中, 一方面可利用注意力池化机制捕捉个体差异贡献的优势, 另一方面也可以利用群体内部组织结构信息. 记第 j 个子群体的行为特征为 g_j^t , 则整个群体的行为特征为

$$G_t = g_j^t \oplus g_{j+1}^t \oplus \cdots \oplus g_m^t, \quad (13)$$

其中, $\oplus$ 表示拼接操作, 整个群体划分为 m 个子群体. 拼接型注意力池化整体结构如图 6 所示.

图 5 (网络版彩图) 多层注意力池化示意图

Figure 5 (Color online) Hierarchical attentive pooling at time-step t

图 6 (网络版彩图) 拼接型注意力池化示意图

Figure 6 (Color online) Attentive pooling with subgroups concatenation

5 实验与结果分析

本节首先介绍了实验使用的数据集、对比模型等设置, 然后讨论网络结构参数的影响并对可视化结果进行分析, 最后列举了使用注意力池化机制模型的性能, 并和当前性能优秀的模型进行对比分析.

5.1 数据集与实验设置

本文选择群体行为识别任务中广泛使用的 The Volleyball Dataset^[8] 上评测所提出的方法. The Volleyball Dataset 包含 55 段排球比赛的视频, 在整体数据集中选取了 4830 个片段中的关键帧进行标

表 1 对比模型

Table 1 Model comparison

Model ID	Model name	Description
B1	HTM-max pooling	Baseline, proposed by [8]
B2	HTM-avg pooling	Replace pooling mechanism in B1 with average pooling
B3	HTM-GAP	Combine the HTM model with GAP
B4	HTM-HAP	Combine the HTM model with HAP
B5	HTM-APSC	Combine the HTM model with APSC

表 2 The Volleyball Dataset 消融实验结果

Table 2 Results of ablation study on The Volleyball Dataset

Type	Method	Accuracy (%)
Integration	B1-HTM-max pooling	70.3
	B2-HTM-avg pooling	68.5
	B3-HTM-GAP(Ours)	74.2
	B4-HTM-HAP(Ours)	77.7
Concatenation	B1-HTM-max pooling	81.9
	B2-HTM-avg pooling	80.7
	B5-HTM-APSC(Ours)	84.5

注, 主要给出了 9 种个体行为与 8 种群体行为的标签以及关键帧中人体在图像坐标系下的坐标. 对于未标注的非关键帧, 本文选择与文献 [8] 一样采用 dlib tracker^[25] 生成相应的人体坐标, 行为类别标签与关键帧相同.

为便于与同类方法进行比较, 本文采用了与之相同的基础模型, 基础模型在 ImageNet^[26] 中完成预训练, 在随后的群体行为识别任务中对 fc6 与 fc7 层的相关参数进行了微调. 个体层级的 LSTM 包含 3000 个节点, 群体层级的 LSTM 包含 2000 个节点, 均利用 xavier 方法^[27] 进行初始化. 整体网络模型采用依次训练个体与群体两层模型的方式, 首先对个体层级 LSTM 网络与 alexnet 网络进行训练, 随后再加入群体层级的 LSTM 一起进行训练. 在模型损失函数中, 超参数 $\lambda = 2.0$, 使用的优化方法为 ADAM^[28], 初始学习率设定为 0.0001, 项目依托 Tensorflow 框架实现.

5.2 实验结果

为验证所提方法的有效性, 本文实验按照不同设置, 选定了 5 个模型进行对比实验, 详细信息如表 1 所示. 在实验中, 本文选择以球网为界, 把整个群体分为左、右侧两个子群体. 从表 2 的实验结果可以发现, 本文所提出的 3 种注意力池化机制在识别性能上均优于原先模型中的池化操作方式, 所提方法行之有效. 在此基础上, 将群体组织结构信息引入全局注意力模型之后, HAP 与 APSC 机制的分类性能进一步得到了提升, 表明两者有效结合可提高群体行为特征判别性. 在排球比赛场景中, 比赛双方运动员分别位于球网的左右两侧, 因此选择上述划分方式.

表 3 中主要展示了使用基于注意力池化的群体行为识别模型与其他先进方法的分类性能对比情况. 实验对比主要在单个群体 (集约型)、两个子群体 (拼接型) 条件分别展开. 在集约型条件下, 本文提出的 HTM-HAP 模型性能优于其他对比方法; 在拼接型条件下, 本文提出的 HTM-APSC 模型取得了本数据集当前领先方法相当的性能.

表 3 The Volleyball Dataset 实验结果对比
Table 3 Comparison of experiment results on The Volleyball Dataset

Type	Method	Accuracy (%)
Integration	B1-Hierarchical LSTM ^[14]	
	CERN (Integration) ^[9]	73.5
	SBGAR ^[10]	66.9
	SRNN (Integration) ^[13]	74.4
	Ours-B3	77.7
Concatenation	B1-Hierarchical LSTM ^[14]	81.9
	CERN (Concatenation) ^[9]	83.3
	SRNN (Concatenation) ^[13]	83.4
	Ours-B5	84.5

表 4 隐层节点数量对模型性能的影响
Table 4 Impact of MLP hidden units in attention module

Hidden layer dim	Method	Accuracy (%)
1024	HTM-GAP	69.0
	HTM-HAP	77.7
	HTM-APSC	83.1
512	HTM-GAP	74.2
	HTM-HAP	77.4
	HTM-APSC	83.4
256	HTM-GAP	73.3
	HTM-HAP	77.4
	HTM-APSC	82.4

5.3 实验讨论

本小节主要从注意力模块中的隐层节点数对分类性能的影响、识别结果可视化与群体行为分类混淆矩阵 3 个方面对实验结果进行讨论分析。

5.3.1 隐层节点数对分类性能的影响

为探究注意力模块中的隐层节点数对分类性能的影响, 本文以隐层节点数为变量展开实验. 表 4 实验结果显示, 在 HTM-GAP 和 HTM-APSC 模型中, 当注意力模块的 MLP 中有 512 个节点时, 其性能较好; 在 HTM-HAP 模型中, 采用 1024 隐层节点数模型性能较优.

5.3.2 识别结果可视化分析

为分析模型作用效果, 定性分析所提出的注意力池化机制, 本文对实验结果进行了可视化, 随机选取了部分实验结果于图 7 进行展示. 图 7 第 1 列表示原视频帧, 第 2 列为识别结果和个体注意力权重, 第 3 列图像根据注意力权重对不同个体周围区域进行亮度调整. 为突出显示个体重要性程度, 本文将每个个体对应的注意力权重数值以人体框左侧红色矩形条的形式加以显示, 矩形条越长表示其重要性程度越高, 相应地对群体行为识别的影响越大. 图中第 1 行所属视频片段的群体行为为“左方二

图 7 (网络版彩图) 实验结果可视化

Figure 7 (Color online) Visualization of experiment results

传”,不同个体对应各自的注意力权重数值,之后各行分别展示了“右方扣球进攻”、“左方一传”等群体行为的实验结果.观察结果可得,本文提出的模型可有效地对群体中的个体重要性进行区分.

5.3.3 群体行为分类混淆矩阵

实验所得群体行为分类混淆矩阵如图 8 所示.从混淆矩阵可发现,“左方本回合获胜”的识别准确率最高,而“右方一传”等群体行为具有较低的识别准确率.当视野内的个体聚集时,本文所提模型识别效果较好,因此对于回合获胜类的群体行为具有较好的识别性能.在排球场景中,“一传”和“二传”具有较高的相似度,“二传”和“扣球”时间间隔非常短,对于视频帧采样和模型学习提出了新的挑战.

6 结论与未来工作

本文提出的注意力池化机制将个体行为按照其对群体行为贡献度的差异性进行区别考量,在个体层级行为特征向群体层级行为特征的过渡过程中自适应地实现了个体行为特征的聚合,有效增强了群体行为特征的判别性与模型分类性能.与本领域之前的方法对比,该机制同时考虑到群体内部的一致性与个体间的差异性,一定程度上解决了个体间上下文关系复杂条件下的特征聚合问题.但是,行为特征的层级提升并不局限于对个体特征进行简单聚合,融合个体-群体关系的图网络相关模型可以更

r_spike	81.6%	3.7%	1.2%	0.9%	0.4%	2.7%	0.1%	9.5%
r_pass	2.9%	79.4%	4.3%	1.5%	1.0%	3.3%	2.5%	5.1%
r_winpoint	0.0%	0.0%	83.7%	1.3%	1.1%	3.6%	5.3%	5.1%
l_set	0.6%	2.7%	0.0%	86.7%	1.8%	7.0%	0.7%	0.6%
l_spike	0.6%	2.5%	0.0%	6.5%	85.9%	3.4%	0.0%	1.1%
l_pass	2.6%	5.5%	2.1%	3.6%	1.0%	83.5%	1.2%	0.4%
l_winpoint	0.0%	0.5%	1.3%	3.5%	1.0%	0.0%	93.7%	0.0%
r_set	3.6%	12.4%	2.1%	0.5%	1.0%	1.8%	0.0%	78.5%
	r_spike	r_pass	r_winpoint	l_set	l_spike	l_pass	l_winpoint	r_set

图 8 实验所得混淆矩阵
Figure 8 Confusion matrix of experiment results

加自然地刻画群体内部蕴含的关联关系, 从而提高群体行为特征的鲁棒性和可解释性. 因此, 未来计划继续深入研究基于图模型提出更有效的群体行为识别方法, 建立新型群体行为分析框架.

参考文献

1 Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos. In: Proceedings of Advances in Neural Information Processing Systems (NeuPS2014), 2014. 568–576

2 Wang L M, Xiong Y J, Wang Z, et al. Temporal segment networks: towards good practices for deep action recognition. In: Proceedings of European Conference on Computer Vision (ECCV), 2016. 20–36

3 Ji S W, Xu W, Yang M, et al. 3D convolutional neural networks for human action recognition. IEEE Trans Pattern Anal Mach Intell, 2013, 35: 221–231

4 Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015. 4489–4497

5 Yu S, Cheng Y, Xie L, et al. Fully convolutional networks for action recognition. IET Comput Vision, 2017, 11: 744–749

6 Varol G, Laptev I, Schmid C. Long-term temporal convolutions for action recognition. IEEE Trans Pattern Anal Mach Intell, 2018, 40: 1510–1517

7 Choi W, Savarese S. A unified framework for multi-target tracking and collective activity recognition. In: Proceedings of European Conference on Computer Vision (ECCV), 2012. 215–230

8 Ibrahim M, Muralidharan S, Deng Z W, et al. A hierarchical deep temporal model for group activity recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 1971–1980

9 Shu T M, Todorovic S, Zhu S C. CERN: confidence-energy recurrent network for group activity recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. 5523–5531

10 Li X, Chuah M C. SBGAR: semantics based group activity recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. 2876–2885

11 Bagautdinov T M, Alahi A, Fleuret F, et al. Social scene understanding: end-to-end multi-person action localization and collective activity recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. 3425–3434

- 12 Wang M S, Ni B B, Yang X K. Recurrent modeling of interaction context for collective activity recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. 3048–3056
- 13 Biswas S, Gall J. Structural recurrent neural network (SRNN) for group activity analysis. In: Proceedings of IEEE Winter Conference on Applications of Computer Vision (WACV), 2018. 1625–1632
- 14 Ibrahim M S, Muralidharan S, Deng Z W, et al. Hierarchical deep temporal models for group activity recognition. 2016. ArXiv:1607.02643
- 15 Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*, 1997, 9: 1735–1780
- 16 Herath S, Harandi M, Porikli F. Going deeper into action recognition: a survey. *Image Vision Comput*, 2017, 60: 4–21
- 17 Chaquet J M, Carmona E J, Fernández-Caballero A. A survey of video datasets for human action and activity recognition. *Comput Vision Image Underst*, 2013, 117: 633–659
- 18 Nabi M, Bue A, Murino V. Temporal poselets for collective activity detection and recognition. In: Proceedings of IEEE Conference on Computer Vision Workshops (ICCV Workshops), 2013. 500–507
- 19 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. In: Proceedings of Advances in Neural Information Processing Systems (NIPS 2017), 2017. 5998–6008
- 20 Mnih V, Heess N, Graves A, et al. Recurrent models of visual attention. In: Proceedings of Advances in Neural Information Processing Systems (NIPS 2014), 2014. 2204–2212
- 21 Sharma S, Kiro R, Salakhutdinov R. Action recognition using visual attention. 2015. ArXiv:1511.04119
- 22 Yang Z C, Yang D Y, Dyer C, et al. Hierarchical attention networks for document classification. In: Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016. 1480–1489
- 23 Long X, Gan C, de Melo G, et al. Attention clusters: purely attention based local feature integration for video classification. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018. 7834–7843
- 24 Lin Z S, Feng M W, Yu M, et al. A structured self-attentive sentence embedding. 2017. ArXiv:1703.03130
- 25 King D E. Dlib-ml: a machine learning toolkit. *J Mach Learn Res*, 2009, 10: 1755–1758
- 26 Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks. In: Proceedings of Advances in Neural Information Processing Systems (NIPS 2012), 2012. 1097–1105
- 27 Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS-2010), 2010. 249–256
- 28 Kingma D, Ba J. Adam: a method for stochastic optimization. 2014. ArXiv:1412.6980

Attentive pooling for group activity recognition

Ding LI^{1,2} & Wensheng ZHANG^{1*}

1. *Institute of Automation, Chinese Academy of Sciences, Beijing 100091, China;*

2. *School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China*

* Corresponding author. E-mail: zhangwenshengia@hotmail.com

Abstract In group activity recognition, the hierarchical framework is widely used to represent the relationships between individuals and their corresponding groups and has achieved promising performance. However, existing methods simply employ the max/average pooling in this framework, overlooking the distinct contributions of different individuals to the group activity recognition. In this paper, we propose a new contextual pooling scheme, named attentive pooling, which enables weighted information transition from individual actions to group activity. Using the attention mechanism, attentive pooling is intrinsically interpretable and can embed the member context in the existing hierarchical model. To verify the effectiveness of the proposed scheme, two specific attentive pooling methods, i.e., global attentive pooling (GAP) and hierarchical attentive pooling (HAP), are designed. GAP rewards individuals significant to the group activity, while HAP further considers the hierarchical division by introducing the subgroup structure. Experimental results on the benchmark dataset demonstrate that the proposed scheme is significantly superior over the baseline and comparable to state-of-the-art methods.

Keywords group activity recognition, representation learning, attention mechanism, deep learning

Ding LI was born in 1993. He received his B.S. degree from Wuhan University and is now a Ph.D. candidate at the Institute of Automation, Chinese Academy of Sciences. His research interests include deep learning, action recognition, and computer vision.

Wensheng ZHANG was born in 1965. He received his Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences in 2001. He is a researcher at the Institute of Automation, Chinese Academy of Sciences. His research interests include statistical machine learning and big data analysis.