文章编号：2095-2163(2021)05-0077-06 中图分类号：TP391 文献标志码：A

基于时空自注意力转换网络的群组行为识别

张天雨，许 飞，江朝晖
（合肥工业大学 计算机与信息学院，合肥 230601）

摘 要：个体间关系信息的获取是群组行为识别中关键问题。为了获取更加丰富的关系信息，本文提出了一种时空自注意力转换网络（Spatio-Temporal Transformer Network）。空间自注意力转换模块可以同时处理群组中的所有个体，包括其外观特征和位置特征，以便提取个体间空间关系信息。使用时序自注意力转换模块进行时序建模。为了获得更加丰富有效的关系信息，提出了全局空间注意图，用以增强模型空间关系推理能力，使用时序掩膜优化时序自注意力转换模块。通过在 Volleyball 和 Collective Activity 数据集上实验验证，结果表明本文方法性能优于其它方法。

关键词：群组行为识别；自注意力转换网络；自注意力

Spatio-temporal transformer network for group activity recognition

ZHANG Tianyu, XU Fei, JIANG Chaohui
(School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230601, China)

【Abstract】The relationship information between individuals is a key problem in group activity recognition. In order to obtain richer relational information, a spatio-temporal transformer network is proposed in this paper. The spatial transformer module processes all individuals in the cluster simultaneously, including their appearance features and location features, to extract information on the spatial relationships between individuals. Then the temporal transformer module is used for temporal sequence modeling. In order to obtain richer and effective relationship information, a global spatial attention map is proposed to enhance the model's spatial relational reasoning ability, and a temporal mask is used to optimize the temporal transformer module. We verify our model on Volleyball and Collective Activity datasets, and the experimental results show that the proposed method achieves advanced performance.

【Key words】group activity recognition; transformer; self-attention

0 引 言

群组行为识别是指对多个个体共同参与的活动进行识别，具有广泛的应用领域。如：体育视频分析、智能视频监控、机器人视觉等。与传统个体行为识别不同的是，群组行为识别需要理解个体之间的交互关系，而个体的位置、行为以及个体之间的交互关系随时间不断变化。

早期的方法使用概率图模型处理手工提取的特征。近几年，循环卷积神经网络（Recurrent Neural Network, RNN）和长短时记忆网络（Long Short-Term Memory，LSTM）凭借其强大的序列信息处理能力，被许多学者用于群组行为识别。Ibrahim M S 等人[1]设计了一个层次 LSTM 模型，其中一个 LSTM 提取成员个体行为动态特征，另一个用于聚合个体层次信息作为场景表示，但在使用 LSTM 聚合个体层次信息时忽略了个体空间关系。Ibrahim M S 等人[2]在之后的工作中引入一个关系层为每个人学习紧凑的关系表示，但这种关系层学习个体关系的方法不够灵活。

为解决上述问题，本文提出时空自注意力转换网络模型用于群组行为识别。首先使用空间自注意力转换模块，灵活地建模个体间的空间关系，其次使用时序自注意力转换模块进行时序建模，最后将时空关系建模后的特征用于群组行为识别。

本文的主要贡献是：提出了一种端到端的时空自注意力转换模型，以及全局空间关注图，改进空间自注意力转换模块；使用时序掩膜策略，优化时序自注意力转换模块。在两个流行数据集上进行验证，均取得了优秀的表现。

1 相关工作

1.1 群组行为识别

早期的研究人员采用概率图模型处理手工提取

作者简介：张天雨（1996-），男，硕士研究生，主要研究方向：计算机视觉；许 飞（1995-），男，硕士研究生，主要研究方向：计算机视觉；江朝晖（1997-），男，硕士研究生，主要研究方向：计算机视觉。
通讯作者：张天雨 Email：tianyuZ061@163.com

收稿日期：2021-02-01

的特征[3-4]。近期,深度学习网络在各个领域取得优异的表现,一些学者将 RNN 以及 LSTM 引入群组行为识别任务中。Ibrahim M S 等人[1]提出了基于 LSTM 的层次模型,将卷积神经网络(Convolutional Neural Network,CNN)和 LSTM 作为骨干网络,其中 LSTM 可以捕捉每个个体的时间动态特征。其后,许多基于 CNN 和 RNN 结合的群体活动识别方法涌现出来。

例如,Shu T 等人[5]提出了能量层和 LSTM 结合的 CERN 网络,能量层用于捕获 CERN 内所有 LSTM 预测之间的依赖关系,并以这种方式通过能量最小化实现更加可靠的识别;Li X 等人[6]使用一个 LSTM 为每个视频帧生成一个标题,另一个 LSTM 根据这些生成的字幕,预测最终的活动类别;Ibrahim M S 等人引入一个关系层模块,该模块可以编码个体与其他个体的关系信息;Tsunoda T 等人[7]设计了一个层次 LSTM,在 LSTM 中引入了保持状态作为一种外部可控状态,并且扩展了分层 LSTM 的集成机制。

此外,一些方法采用注意力机制来确定与群组活动中的关键人物。例如 Ramanathan V 等人[8]结合双向长短时记忆网络(Bi-directional Long Short-Term Memory,BLSTM)和注意力(Attention)机制,提出注意力模型,给予事件中关键参与者更高的权重。Qi M 等人[9]还利用注意机制同时从视觉域和语义域寻找关键人物。本文基于自注意力机制,提出时空自注意力转换网络进行群组行为识别。

1.2 自注意力机制(Self-Attention)

自注意力机制是自注意力转换网络(Transformer)的基础模块[10],用于为序列的所有实体之间的交互建模,在自然语言处理领域表现优异。原理上,自注意力层通过聚合来自完整输入序列的全局信息,来更新序列的每个组成部分。其输入由一组查询(Queries,Q)、维度为 D 的键(Keys,K),和值(Values,V)组成,将这些输入打包成矩阵形式实现高效计算。首先将 Q 与 K 的转置矩阵相乘并除以 $\sqrt{D}$,再使用 softmax 层进行归一化,以获得注意力分数。序列中每个实体更新为序列中所有实体的加权和,其中的权重由注意力分数给出。其公式为:

$$SelfAtttention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{D}}\right)V. \quad (1)$$

这种自注意力机制被用于许多关系建模、目标检测等计算机视觉任务。本文工作中利用基于自注意力机制的 Transformer 用于时空关系建模。

2 模型框架

2.1 总体框架

网络由个体特征提取、基于 Transformer 的时空特征融合模块和残差连接特征融合模块 3 部分组成。网络框架如图 1 所示。网络输入为视频帧序列 $F = \{F_1, F_2, \cdots, F_T\}$ 以及个体边界框 B;使用 2D CNN 网络提取输入视频帧的特征图;RoiAlign 层[11]根据个体边界框 B 提取个体外观特征;使用 FC 层将每个个体成员特征映射为维度 1×1 024,将其称为原始个体特征;将提取的个体特征输入时空 Transformer 模块,进行时空信息建模。为了减少深度网络退化问题,采用残差链接将原始特征与时空信息建模后的成员特征融合,最后使用分类层进行分类。

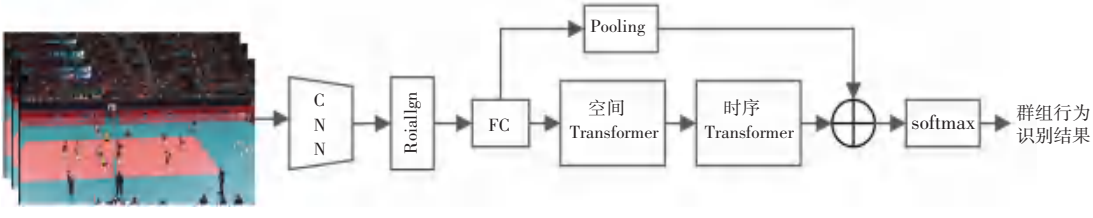

图 1 时空自注意力转换网络结构

Fig. 1 Structure of Spatio-Temporal Transformer Network

2.2 空间 Transformer

在将原始特征输入该模块之前,需根据个体边界框为原始特征添加空间位置信息。对于个体 i,根据其边界框中心点 (x_i, y_i),使用 Vaswani A 等人[13]提出的 PE 位置编码函数对其进行编码。编码得到的空间位置信息维度和个体 i 的特征维度相同,其前一半维度为 x_i 的编码,后一半为 y_i 的编码。编码函数为:

$$PE(pos, 2i) = sin(pos/10000^{2i/D_{in}}), \quad (2)$$

$$PE(pos, 2i+1) = cos(pos/10000^{2i/D_{in}}), \quad (3)$$

其中,pos 为个体的位置;i 为空间位置编码向量的维度;D_{in} 的值等于个体特征维度大小的一半。

空间位置信息x_i和y_i均使用上述编码，编码后将其使用 concatenate 方式进行连接。空间位置编码与个体特征具有相同的维度，将两者相加得到具有空间位置信息的个体特征。

空间 Transformer 原理如图 2 所示。Transformer 由 L 层组成，每层有 2 个子层：一个多头注意层和一个前馈层。其原始输入为经空间位置信息编码后的特征矩阵 $\boldsymbol{X} \in R^{N\times D}$。其中，$N$ 代表节点数量，D 表示通道数。对于 H 个注意头的第 j 个头的注意层，计算其输出 $X_{j\in} R^{N\times d}, d = D/H$：

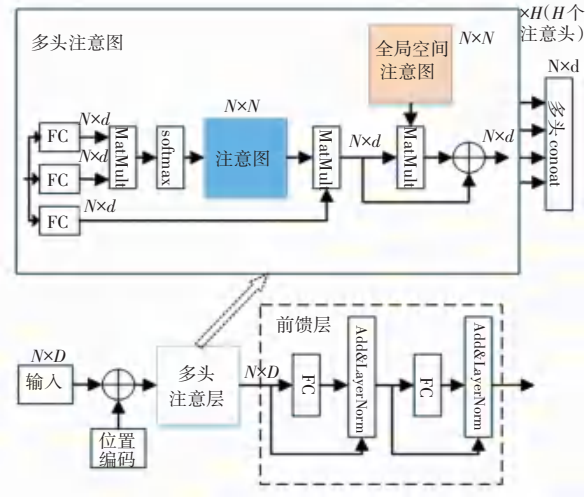

图 2　空间 Transformer 原理

Fig. 2　Principle of Spatio-Transformer Network

$$X_j = softmax\left(\frac{Q_j K_j^T}{\sqrt{d}}\right)V_j, Q_j = X W_j^q, K_j = X W_j^k, V_j = X W_j^v, \quad (4)$$

其中，$\boldsymbol{W}_j^q$、$\boldsymbol{W}_j^k$ 和 $\boldsymbol{W}_j^v$ 维度均为 $R^{N\times d}$，3 个权重矩阵将输入 X 映射为查询、键和值。对每个注意头重复此操作，将多个头的输出，使用 concatenation 方式连接。再使用一个全连接层（权重矩阵 $\boldsymbol{W}^{l_1} \in R^{D\times D}$），将上一层输出映射为维度 $R^{N\times D}$ 的特征矩阵，由两个全连接层组成前馈层（权重矩阵为 $\boldsymbol{W}^{l_2} \in R^{D\times D_1}$，$\boldsymbol{W}^{l_3} \in R^{D_1\times D}$）。多头注意层和前馈层的输出都使用残差函数，并且对残差连接融合的特征进行层归一化。其中激活函数 σ 为 ReLU，最终得到的输出 X_o 和输入 X 维度相同：

$$X_a = LayerNorm(X + Dropout(concat(X_1, X_2, \cdots, X_H) W^{l_1})), \quad (5)$$

$$X_o = LayerNorm(X_a + Dropout(\sigma(X_a W^{l_2}) W^{l_3})). \quad (6)$$

2.3　全局空间注意图

如上所述，使用空间 Transformer 中多头注意力为每个时刻个体计算空间关注度，这是一种随时间变化的空间关注度。由于每个个体在群组活动中扮演特定的角色，可在整个群组活动过程中设定一个时序共享的全局空间注意力模块，来强制模型学习更多不同时刻的一般关注。

如图 2 所示，在多头注意力和前馈层之间加入 K 全局注意图 $A_{global} = \{A_{global}^1, A_{global}^2, \cdots, A_{global}^K\}$，$A_{global}^k \in R^{N\times N}$，在这里 K 取 N 值。所有数据样本共享全局注意图，代表整个群组活动内在关系模式，多个图构成全局注意图增加网络泛化能力。本文将其作为网络的参数，并与模型一起进行优化。该模块结构简单、参数少，但消融实验表明其效果显著。全局空间关注度模块使用残差函数，增加该模块后计算公式表示为：

$$X_{mid} = LayerNorm(X_a + Dropout(\bar{A}_{global} X_a)). \quad (7)$$

其中，$\bar{A}_{global}$ 为 K 个图相加求平均，X_a 为 Transformer 中多头注意力的输出，X_{mid} 为全局空间关注度模块的输出，并作为前馈层的输入。

2.4　时序 Transformer

时序 Transformer 和空间 Transformer 具有相同的原理，其不同之处在于输入特征为时序特征以及多头注意层的计算方式。输入的时序特征由各时刻空间特征在个体维度最大池化获得。在时序特征经多头注意层时，对多头注意层中计算出的关注图矩阵后，增加一个掩膜矩阵 $\boldsymbol{M} \in R^{N\times N}$。$\boldsymbol{M}$ 矩阵为：

$$\boldsymbol{M}(m_1, m_2) = \begin{cases} 1, & |m_1 - m_2| \leqslant \gamma, \\ 0, & \text{other.} \end{cases} \quad (8)$$

其中，m_1、m_2 为矩阵的行和列，γ 为时间窗口大小，设置为输入单个视频序列帧数的一半。增加掩膜后的注意力层计算为：

$$MaskAtttention(Q,K,V) = softmax\left(\frac{Q K^T}{\sqrt{D}} \circ M\right)V. \quad (9)$$

其中，$\circ$ 表示 Hadamard 乘积。因此，当为某个时序特征进行时序建模时，只考虑该时刻前后 γ 时刻内的时序特征，其它时刻的注意分数被设为零。采用这种策略，减少了时序建模时信息冗余，降低了时序建模难度。

2.5　损失函数

将时序 Transformer 的输出与原始特征进行求和融合，形成最终场景表示，将场景表示送入分类层进行群组行为识别。使用空间 Transformer 的输出特征与原始特征求和融合后计算个体损失。整个模型以反向传播端到端方式训练，损失函数由个体损失和群组损失组成，其公式如下：

$$Loss = L(y^G, y_{gt}^G) + L(y^P, y_{gt}^P). \quad (10)$$

其中，L 为交叉熵损失函数；y_{gt}^G 和 y_{gt}^P 是群组行为和个体行为标签；y^G 和 y^P 是预测值。

3 实验结果与分析

3.1 数据集

(1) Volleyball 数据集。数据集由 55 个排球比赛视频中截取的 4 830 个视频片段组成[1]。每个视频片段中间帧标注了个体边界框、个体行为标签以及群组行为标签。其中个体行为标签有 9 种，群组行为标签共有 8 种。对于每个带标注的帧，该帧周围有多个未带标注的帧可用。实验中使用一个长度为 $T = 10$ 的时间窗口，对应于标注帧的前 5 帧和后 4 帧。未被标注的个体边界框数据从该数据集提供的轨迹信息数据获取。使用 3 494 个视频片段作为训练集，1 337 个视频片段作为测试集。

(2) Collective Activity 数据集。数据集由低分辨率相机拍摄的 44 个视频片段组成，总共约 2500 帧[3]。每个视频片段每 10 帧有一个标注，标注包含个体行为和群组行为标签，以及个体的边界框。共 5 个群组活动标签，6 个个体行为标签。实验中 2/3 视频用于训练，其余的用于测试。

3.2 实验细节及评价标准

对于 Volleyball 数据集，网络超参设置如下：最小批量大小为 8，Dropout 参数为 0.3，学习率初始设置为 1E-4，网络训练 180 个周期，每 30 个周期学习率将为之前的 0.5 倍，学习率在 4 次衰减后停止衰减。空间自注意力转换模块层数为 1，注意头数为 2，时序自注意力转换模块层数和注意头数均为 1。实验采用 ADAM(ADAptive Moment)优化器。

在 Collective Activity 数据集上，网络超参设置为：最小批数据大小为 16，Dropout 参数为 0.5，初始学习率为 1E-3，每 10 个周期学习率将为之前的 0.1 倍，学习率在四次衰减后停止衰减。网络共训练 80 个周期。空间自注意力转换模块层数为 1，注意头数为 2，时序自注意力转换模块层数和注意头数均为 1。实验采用 ADAM 优化器。

3.3 消融实验

3.3.1 基线模型设计

为通过消融实验来证明本文模型中各个模块的有效性，设计以下变体模型：

B1(Baseline)：基于个体特征模型。在该模型中，采用 Inception-v3 来计算每个帧中个体的高维特征。将这些特征经平均池化，计算出群组行为的特征。这些特征被送到 Softmax 分类器中，以预测每个帧中群组行为的标签。视频的预测标签为所有视频帧的预测标签，通过求和平均得到。

B2(Baseline + ST)：该变体模型使用空间 Transformer(Spatio-Transformer, ST)对 Inception-v3 提取的个体特征进行空间关系推理。

B3(Baseline+ST+TT)：该变体在 B2 的基础上增加无掩膜优化的时序(Temporal - Transformer, TT)，对时序关系进行推理。

B4(Baseline+ST_Enhance+TT)：在 B3 的基础上增加全局空间注意图增强，对空间关系的推理。

B5(Baseline+ST_Enhance+TT_Enhance)：为本文的最优模型，在 B4 的基础上增加掩膜对 TT 进行优化。

3.3.2 实验结果分析

模型及其变体在 Volleyball 数据集上的识别准确率结果见表 1。本文提出的 B5 模型取得了最好的性能。其达到 92.52%的最高准确率，与基线模型 B1 相比准确率提升了 3.37%。与 B1 相比，变体模型 B2 通过探索个体之间的空间交互，识别准确率提高了 0.87%。B3 被用来说明在时间和空间领域捕捉个体空间交互关系以及时序关系的重要性，B4 和 B3 相比提高了 0.9%的准确率，证明了全局空间注意图这种不同时刻的一般关注对于识别群体活动的有效性。B5 和 B4 相比，验证了通过增加 MASK 减少在时序关系推理时的信息冗余，可以提高模型的性能。

表 1 Volleyball 数据集上的消融实验结果

Tab. 1 Ablation results on Volleyball dataset

消融因素	Accuracy(单位:%)
B1:Baseline	89.15
B2:Baseline+ST	90.02
B3:Baseline+ST+TT	91.40
B4:Baseline+ST_Enhance+TT	92.30
B5:Baseline+ST_Enhance+TT_Enhance	92.52

3.4 与各方法的对比分析

表 2 显示了本文的最佳模型与各方法在 Volleyball 数据集上的比较结果。由表 2 可知，本文方法在 Volleyball 数据集上达到最好的表现。和 HRN 模型相比，虽然其模型包括个体之间的关系信息，但其方法提取空间关系未充分利用空间信息。因此，本文模型优于 HRN 模型。和 ARG 模型相比，虽然该模型充分探究了个体间空间位置和外观关系，但在时序建模方面采用时序抽样策略没有完整

利用时序信息，而本文模型采用了时序关系建模优化，因此本文模型优于 ARG 模型。

表 2　各方法在 Volleyball 数据集上的准确率

Tab. 2　Accuracies of different methods on Volleyball dataset

Method	Accuracy(%)
HDTM[1]	81.90
CERN[5]	83.30
SBGAR[6]	66.90
stagNet[9]	89.30
SSU[13]	90.60
HRN[2]	89.50
ARG[14]	92.50
Ours	92.52

在 Collective Activity 数据集上与其它先进方法进一步比较结果见表 3。本文模型表现优于其它方法，达到 91.24%的群体活动识别准确率。结果表明了该模型捕获时空关系信息的有效性和通用性。

表 3　各方法在 Collective Activity 数据集上的准确率

Tab. 3　Accuracies of different methods on Collective Activity dataset

Method	Accuracy(%)
HDTM[1]	81.50
CERN[5]	87.20
stagNet[9]	89.10
ARG[14]	91.00
Ours	91.24

3.5　数据可视化

(1)空间注意力可视化。在图 3 中可视化了本文模型在 Volleyball 数据集上两个空间注意头生成注意力图的例子。根据注意力图，在图像中使用红星标出了关键个体。可视化结果表明本文模型能够捕捉群体活动中关键关系信息。

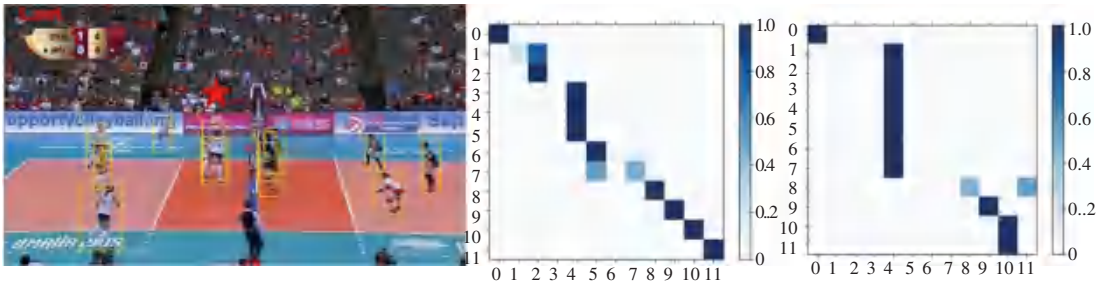

图 3　空间注意力可视化

Fig. 3　Spatial attention visualization

(2)t-SNE 可视化。图 4 显示了 t-SNE 可视化不同模型变体在 Volleyball 数据集上学习的视频表示。使用 t-SNE 将排球数据集的验证集上的视频表示投射到二维空间。从图上可以观察到，本文的 B5 模型学习的群组场景表示具有较好的分离度，且全局空间注意力增强和时序掩膜优化结合，可以更好地区分群体活动。

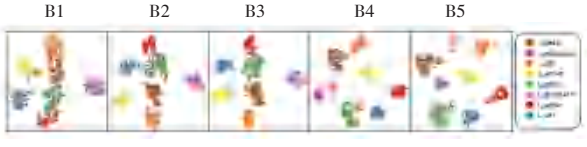

图 4　不同变体模型视频表示的 t-SNE 可视化

Fig. 4　t-SNE visualization of video representations of different variants of the model

4　结束语

本文提出一种灵活有效的方法对群组中个体进行时空关系推理，基于自注意力机制的时空 Transformer 关系网络获得用于群组行为识别的视频表示。在当前流行数据集上的实验表明，本文方法和当前优秀方法相比准确率更高。并可视化了部分网络，可以更加了解网络的工作原理。

参考文献

[1] IBRAHIM M S, MURALIDHARAN S, DENG Z, et al. A hierarchical deep temporal model for group activity recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 1971-1980.

[2] IBRAHIM M S, MORI G. Hierarchical relational networks for group activity recognition and retrieval[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 721-736.

[3] CHOI W, SHAHID K, SAVARESE S. What are they doing?: Collective activity classification using spatio-temporal relationship among people[C]//2009 IEEE 12th international conference on computer vision workshops, ICCV Workshops. IEEE, 2009: 1282-1289.

[4] CHOI W, SHAHID K, SAVARESE S. Learning context for collective activity recognition[C]//CVPR 2011. IEEE, 2011: 3273-3280.

[5] SHU T, TODOROVIC S, ZHU S C. CERN: confidence-energy recurrent network for group activity recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 5523-5531.

(下转第 87 页)