Journal of System Simulation

Volume 36 | Issue 1 Article 5

1-20-2024

Action Recognition Model of Directed Attention Based on Cosine Similarity

Chen Li
Command & Control Engineering College, Army Engineering University of PLA, Nanjing 210007, China, 651220007@qq.com

Ming He
Command & Control Engineering College, Army Engineering University of PLA, Nanjing 210007, China, paper_review@126.com

Chen Dong
Military Human Resource Support Center, Political Work Department of the Army, Beijing 100072, China

Wei Li
Command & Control Engineering College, Army Engineering University of PLA, Nanjing 210007, China

Action Recognition Model of Directed Attention Based on Cosine Similarity

Abstract

Abstract: Aiming at the lack of directionality of traditional dot product attention, this paper proposes a directed attention model (DAM) based on cosine similarity. To effectively represent the direction relationship between the spatial and temporal features of video frames, the paper defines the relationship function in the attention mechanism using the cosine similarity theory, which can remove the absolute value of the relationship between features. To reduce the computational burden of the attention mechanism, the operation is decomposed from two dimensions of time and space. The computational complexity is further optimized by combining linear attention operation. The experiment is divided into two stages : Four ablation experiments are carried out on each module of directed attention to show the best performance of DAM in accuracy and efficiency; the accuracy of the model is 7.3% higher than that of I3D-NL on the Sth-Sth V1(something something V1) dataset and 95.7% on the UCF101(101 human action classes from videos in the wild) dataset. The research results have a wide application prospect in safety monitoring, automatic driving, and so on.

Keywords

action recognition, deep learning, attentional mechanism, cosine similarity, time-space decomposition

第 36 卷第 1 期 系统仿真学报© Vol. 36 No. 1
2024 年 1 月 Journal of System Simulation Jan. 2024

基于余弦相似性的定向注意力行为识别模型

李晨[1]，何明[1*]，董晨[2]，李伟[1]
(1. 陆军工程大学 指挥控制工程学院，江苏 南京 210007；2. 陆军政治工作部 军事人力资源保障中心，北京 100072)

摘要：针对传统点积注意力缺乏方向性的问题，建立了一种基于余弦相似性的定向注意力模型(directed attention model，DAM)。*为有效表示视频帧时空特征间的方向关系，运用余弦相似性理论，定义了注意力机制中关系函数，能够去除特征间关系绝对值；为降低注意力机制计算量，从时间和空间两个维度上对运算进行分解；结合线性注意力运算，进一步优化计算复杂度。*实验分为两个阶段：对定向注意力各模块开展了4个消融实验，以表现DAM在精确度和效率方面的最佳性能；该模型在Sth-Sth V1(something something V1)数据集上的精确度较I3D-NL(inflated 3D ConvNet non-local)高7.3%，在UCF101(101 human action classes from videos in the wild)数据集上的识别精确率为95.7%。研究成果在安全监控、自动驾驶等方面应用前景广泛。

关键词：行为识别；深度学习；注意力机制；余弦相似性；时空分解
中图分类号：TP391.4 文献标志码：A 文章编号：1004-731X(2024)01-0067-16
DOI: 10.16182/j.issn1004731x.joss.22-0937
引用格式: 李晨, 何明, 董晨, 等. 基于余弦相似性的定向注意力行为识别模型[J]. 系统仿真学报, 2024, 36(1): 67-82.
Reference format: Li Chen, He Ming, Dong Chen, et al. Action Recognition Model of Directed Attention Based on Cosine Similarity[J]. Journal of System Simulation, 2024, 36(1): 67-82.

Action Recognition Model of Directed Attention Based on Cosine Similarity

Li Chen[1], *He Ming*[1*], *Dong Chen*[2], *Li Wei*[1]
(1. Command & Control Engineering College, Army Engineering University of PLA, Nanjing 210007, China;
2. Military Human Resource Support Center, Political Work Department of the Army, Beijing 100072, China)

Abstract: Aiming at the lack of directionality of traditional dot product attention, this paper proposes a directed attention model (DAM) based on cosine similarity. *To effectively represent the direction relationship between the spatial and temporal features of video frames, the paper defines the relationship function in the attention mechanism using the cosine similarity theory, which can remove the absolute value of the relationship between features. To reduce the computational burden of the attention mechanism, the operation is decomposed from two dimensions of time and space. The computational complexity is further optimized by combining linear attention operation.* The experiment is divided into two stages : Four ablation experiments are carried out on each module of directed attention to show the best performance of DAM in accuracy and efficiency; the accuracy of the model is 7.3% higher than that of I3D-NL on the Sth-Sth V1(something something V1) dataset and 95.7% on the UCF101(101 human action classes from videos in the wild) dataset. The research results have a wide application prospect in safety monitoring, automatic driving, and so on.

Keywords: action recognition; deep learning; attentional mechanism; cosine similarity; time-space decomposition

收稿日期：2022-08-09 修回日期：2022-10-18
基金项目：江苏省重点研发计划(BE20200729，BE2019762，BE2021729)；军内科研项目(LJ20212Z010032)；军队重点课题(JYKYA2021029)
第一作者：李晨(1994-)，男，硕士生，研究方向为计算机视觉。E-mail：651220007@qq.com
通讯作者：何明(1978-)，男，教授，博导，博士，研究方向为智联网、无人指控等。E-mail：paper_review@126.com

0 引言

当前，AR(action recognition)逐渐成为计算机视觉领域的热门研究课题，具备着广泛应用前景，例如识别作业人员行为，增强安全生产场所防护能力。AR不仅要关注帧内空间特征，还需注意帧间时间特征，以有效提取视频中行为时空特征，实现精确识别。虽然NL(non-local)[1]利用自注意力机制提取全局特征，但其通过点积乘法得到的特征相似关系缺乏方向性，而AR任务较依赖视频帧的时空顺序，例如同一组视频帧若时间排序不同则会呈现不同的行为。另外对于时空位置较多的特征图，NL会耗费较大计算量，这显然不利于即插即用于CNN(convolutional neural network)。因此，如何构造出能够提取方向关系特征的高效注意力行为识别模型，是待需解决的问题。

本文基于余弦相似度[2]提出了DAM(directed attention model)。DAM的核心是DA-NL(directed attention non-local)，其使用余弦相似性提取特征间的方向性全局关系。另外通过时空分解方式将DA-NL运算进行分离以提高计算效率，进而结合Poly-NL(polynomials non-local)[3]和NL[1]设计了6种具备定向注意力的多样性全局模块，使其在低于原始计算代价的情况下，实现时空交互，并通过对各模块的消融实验进一步分析了各模块性能表征。最终基于各定向注意力模块构造的DAM与最近的AR模型在精确度、计算量、参数量方面进行了对比，DAM较NL在总计算成本上具备更高的效率，并在定向注意力的加持下较NL具备更高的识别精确度。

1 相关工作

AR发展初期，以改进密集轨迹(improved dense trajectories, iDT)[4]为代表的传统人工方法占据主导。但随着CNN发展，深度学习方法逐渐成为主流，AR深度学习方法可从时间特征提取、高效率优化和全局特征捕获3个角度概括，如表1所示。

表1 AR方法总览概况
Table 1 Overview of AR methods

角度	方法	代表模型	优、缺点
时间特征提取	双流网络	Two-stream Network[5-8]	可提取时间特征，但计算量高、不稳定、特征局部
	3D CNN	C3D[9]、I3D[10]	可提取时空特征，但计算量高，无法提取全局特征
	时态模块	TSM[11]、TRN[12]、NLNN[1]	可灵活提取时空特征，但无法高效联系全局信息
高效率优化	输入数据优化	TSN[13]	可降低效率，但影响识别准确性
	时空分解卷积	P3D[14]、S3D[15]	可降低效率，但不利于模型最优迭代
	深度分离卷积	CSN[16]	可降低效率，但缺少跨通道信息
	混合卷积	ECO[17]、X3D[18]	可降低效率，但前期训练工作较困难
全局特征捕获	全局均匀采样	TSN[13]	可捕获时间全局特征，但缺乏对空间全局特征
	LSTM	I3D-LSTM[19]	增强全局表征，但训练效率较低
	自注意力机制	NLNN[1]、DirectFormer[2]	可提取全局特征，但缺乏特征方向性和运行效率

(1) 时间特征提取

相较图像识别，AR不仅要关注空间特征，还要聚焦时间特征理解运动信息。具体包含以下3个方面：①基于光流的双流网络[5-8]。两条2D CNN路径以视频帧和堆叠光流图为输入分别提取空间及时间特征，光流表征图像变化携带运动信息，可有效提取时间特征，实现了与iDT比拟的识别效果。但是易受光线变化影响，对存储量和计算量要求较高，且小位移特性不利于识别全局特征。②基于三维卷积的3D CNN[9-10, 20]。其通过将二维卷积扩展为三维卷积，直接提取行为时空特征。如膨胀三维卷积网络(inflated 3D ConvNet, I3D)[10]将经过图像数据集预训练的2D CNN中的二维卷积核参数沿时间轴复制，得到初始化的3D

DOI: 10.16182/j.issn1004731x.joss.22-0937

CNN，为AR使用图像识别成熟架构提供了便利。但3D CNN参数量和计算量较高，并且三维卷积的局部性，导致无法实现对全局特征的有效提取。③基于时态建模的时态模块[1, 11-12, 21-23]。其不需光流和3D卷积扩展时间维度，可灵活嵌入至各2D CNN，具备有效的时空特征提取能力。例如，时间位移模块(temporal shift module, TSM)[11]通过移位操作高效提取帧间信息关系，NL[1]利用自注意力机制交互时空位置间的特征关系。但TSM只能实现局部帧间特征关系的提取、NL计算量受限于时空位置数量，不利于模块灵活嵌入。在时间特征提取方面，本文DAM聚焦于时态模块方法，设计了灵活的DA-NL模块嵌于ResNet50-C2D (residual network 50-convolution 2D)实现行为识别。

(2) 高效率优化

AR作为图像识别的时序扩展，计算量极大，双流网络中光流需额外存储和计算，3D CNN计算量更为昂贵，基于自注意力机制的时态模块同样需耗费较大计算量。具体包含以下3个方面：①输入数据优化。时间段网络(temporal segment networks, TSN)[13]对视频帧均匀采样，降低信息冗余，以较低代价实现学习。隐式双流网络[24]将能够从视频帧中产生类光流的移动网络与时间流CNN连接，缓解了光流计算开销。但从数据端进行效率优化，会损失特性信息，影响识别准确性。②时空分解卷积。其将3D卷积分解，近似为空间2D卷积和时间1D卷积的外积。例如，伪3D(pseudo 3D, P3D)[14]使用1×3×3和3×1×1卷积模拟3×3×3的3D卷积，参数量得到显著降低，且可利用2D CNN初始化训练。分离3D(separable 3D, S3D)[15]还采用Top-heavy方式简化特征量优化效率。但是，硬性的时空分解不利于模型最优迭代，影响识别精细度。③深度分离卷积。如通道分离卷积网络(channel separated convolutional networks, CSN)[16]，其将$T\times H\times W\times C$的3D卷积分解为$T\times H\times W\times 1$的逐通道卷积和$1\times 1\times 1\times C$的逐点卷积，两个部分可在Bottleneck结构基础上共同作用优化模型效率。但其逐通道卷积缺少跨通道信息，缺乏空间关联，影响时空特征提取。④混合卷积。例如，高效卷积网络(efficient convolutional network for online, ECO)[17]通过2D CNN得到特征图后串联3D CNN实现分类、快慢网络(SlowFast)[25]并联配置2D和3D卷积的双分支提取空间和时间特征，扩展3D(X3D)[18]在深度、空间、时间、通道、采样等指标上自动逐步扩展反馈评估以生成高效CNN。本文DAM聚焦于时空分解和混合思想，将DA-NL的注意力机制进行时空分解，并混合了线性注意力降低计算量，缓解了自注意力机制受限于时空位置数量的问题，更利于模块灵活嵌入。

(3) 全局特征捕获

基于卷积操作的CNN缺乏对全局特征捕获的能力，而AR涉及范围更广的时空特征。具体包含以下3个方面：①全局均匀采样。如TSN[13]和TDN(temporal difference networks)[26]通过稀疏采样得到时序上的全局采样帧，实现全局特征提捕获，但此策略缺乏对空间特征的全局把握。②长短时记忆(long short term memory, LSTM)。例如，时间分段LSTM(temporal segment LSTM, TS-LSTM)[27]和I3D-LSTM[19]首先使用2D CNN提取空间特征，再输入LSTM进行融合实现时序特征提取，增强全局表征能力。但LSTM时序先后顺序的严格迭代比较影响训练效率。③自注意力机制。其能直接关注到任意时空位置间的全局信息，在运算上具备很强的并行性。基于点积注意力机制的NL[1]可灵活嵌于CNN捕获全局特征，视频Transformer网络(video transformer network, VTN)[28]和联合Transformer(unified transformer, UniFormer)[29]将CNN与Transformer串联关注全局信息，但点积注意力的特征权值会削减特征间方向性，不利于行为的方向特征提取。视频视觉Transformer(video vision transformer, ViViT)等[2, 30-33]完全摈弃CNN，使用纯Transformer进行AR任务，具备极强的全局特征捕获能力。但局部归纳性能偏弱，不具备

CNN的平移不变性和局部性，需要大量数据才能泛化到AR任务上。本文聚焦于自注意力机制，设计了基于余弦相似性的定向非局部模块(DA-NL)以关注特征间的方向关系，并且可与CNN灵活组合实现快速拟合。

综上所述，本文首先聚焦于时态模块方法，基于NL设计了灵活的DA-NL模块嵌于ResNet50-C2D实现行为识别；其次聚焦于时空分解和混合思想，将DA-NL的注意力机制进行时空分解，并混合了线性注意力降低计算量，缓解了自注意力机制受限于时空位置数量的问题，更利于模块灵活嵌入；最后聚焦于自注意力机制，设计了基于余弦相似性的定向非局部模块(DA-NL)以关注特征间的方向关系，可与CNN灵活组合实现快速拟合。

2 定向注意力设计

本节首先阐述了基于余弦相似性[2]的DA-NL，其次对注意力进行时空分解，进而结合NL、Poly-NL优化定向注意力效率。

2.1 定向注意力模块

NL虽然通过点积乘法得到了各时空位置间的相似性权重，并聚合各特征重新校准了每个时空位置特征，但其相似度权重包含的特征绝对值会对相似性功能造成影响。因为在绝对值较大的情况下，即使特征向量间差异较大(即向量间夹角较大)，也会出现相似度权重较大情况，从而导致特征向量较相似的假象。所以，绝对值造成的相似性作弊现象会减弱各位置特征间的方向性，并通过聚合最终影响输出的注意力特征矩阵。

因此，本文基于余弦相似度计算特征向量间的方向性相似关系，提出DA-NL提取定向注意力全局特征。因相似特征之间的角度应该更小，而余弦相似度可以作为一种正则化机制去除向量幅度，增加特征向量幅度的动机较小，学习的权重总体上也更小，所以其更利于衡量特征间的方向相似度，能更直接地最大化相似特征。

给定一个输入特征张量 $\mathcal{X}\in\mathbb{R}^{T\times H\times W\times C}$，将其折叠为 $\boldsymbol{X}\in\mathbb{R}^{N\times C}$ 作为DA-NL的输入特征矩阵，其中 $N=T\times H\times W$ 即时空维度合并到 N。如等式(1)所示，$\boldsymbol{X}$ 分别转换为查询特征矩阵 $\boldsymbol{Q}\in\mathbb{R}^{N\times C_q}$、关键特征矩阵 $\boldsymbol{K}\in\mathbb{R}^{N\times C_k}$、值特征矩阵 $\boldsymbol{V}\in\mathbb{R}^{N\times C_v}$。

$$\begin{aligned}\boldsymbol{Q}&=\boldsymbol{X}\boldsymbol{W}_\theta\\ \boldsymbol{K}&=\boldsymbol{X}\boldsymbol{W}_\phi\\ \boldsymbol{V}&=\boldsymbol{X}\boldsymbol{W}_g\end{aligned}\tag{1}$$

式中：$\boldsymbol{W}_\theta\in\mathbb{R}^{C\times C_q},\boldsymbol{W}_\phi\in\mathbb{R}^{C\times C_k},\boldsymbol{W}_g\in\mathbb{R}^{C\times C_v}$ 是3个可学习参数线性矩阵。$\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 具有相等特征维度即 $C_q=C_k$，但两者互为不同以避免相似性对称。$\boldsymbol{V}$ 中每个时空位置以相似度矩阵为权值，对所有位置的值特征进行加权求和，得到输出特征。

基于余弦相似性[2]计算 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 的相似性度矩阵，如式(2)和(3)所示：

$$\cos\left(\boldsymbol{Q}_i,\boldsymbol{K}_j\right)=\frac{\boldsymbol{Q}_i\boldsymbol{K}_j}{|\boldsymbol{Q}_i|\times|\boldsymbol{K}_j|}=\cos\theta_{ij}\tag{2}$$

$$\cos\theta_{ij}=\frac{\sum_{c=1}^{C}Q_{i,c}\times K_{j,c}}{\sqrt{\sum_{c=1}^{C}\left(Q_{i,c}\right)^2}\times\sqrt{\sum_{c=1}^{C}\left(K_{j,c}\right)^2}}\tag{3}$$

式中：$\boldsymbol{Q}_i$ 和 $\boldsymbol{K}_j$ 分别为 i 和 j 时空位置的特征向量；θ_{ij} 为 $\boldsymbol{Q}_i$ 和 $\boldsymbol{K}_j$ 的夹角；$Q_{i,c}$ 和 $K_{j,c}$ 分别为 $N=i$、$C=c$ 和 $N=j$、$C=c$ 的向量特征值，C 表示通道数。

由式(2)和(3)可得，余弦相似性去除了特征绝对值 $|\boldsymbol{Q}_i|\times|\boldsymbol{K}_j|$ 的影响，由 $\cos\theta_{ij}$ 独立实现。进而，基于余弦相似性权重，对所有位置的值特征向量加权求和：

$$\boldsymbol{Y}_i=\sum_{j=1}^{N}\cos(\boldsymbol{Q}_i,\boldsymbol{K}_j)\times\boldsymbol{V}_j=\sum_{j=1}^{N}\cos\theta_{ij}\times\boldsymbol{V}_j\tag{4}$$

式中：$\boldsymbol{V}_j$ 为 j 时空位置的值特征向量；$\boldsymbol{Y}_i$ 为 i 时空位置的注意力特征向量。由式(4)可得，$\boldsymbol{Y}_i$ 是所有位置 $\boldsymbol{V}_j$ 的求和，$\boldsymbol{V}_j$ 的权重是 $\cos\theta_{ij}$ 值。$\boldsymbol{Y}_i$ 聚合了全部位置的值特征向量，各位置值特征向量的权重是对应位置键值特征向量和 i 位置查询特征向量的夹角余弦值 $\cos\theta_{ij}$。因此，最终输出的注意力特征矩阵中，每个 i 位置上的特征向量都聚合了所有位置特征向量在 i 位置特征向量方向上的投影。这

DOI: 10.16182/j.issn1004731x.joss.22-0937

种方式可以避免绝对值影响各位置向量的相似性，更有助于cos值的发挥，从而增强自注意力机制的方向性。

为了使定向注意力机制能够即插即用在CNN中，构造了2种定向注意力模块：DA-NL-1(如图1所示，符号‘$\boldsymbol{C}$’表示余弦相似性运算，‘×’表示点积运算，灰色矩形表示核大小为1的卷积和归一化函数)和DA-NL-2。

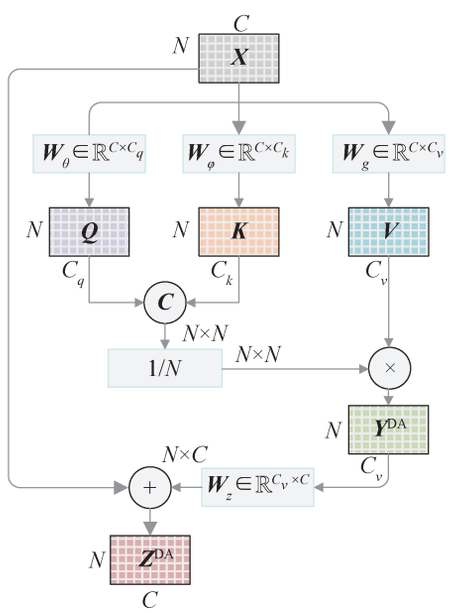


图1 DA-NL-1定向注意力非局部块

Fig. 1 Non-local module of directed attention DA-NL-1

(1) DA-NL-1：使用余弦相似性函数cos，归一化函数$U(x)$如式(5)所示：

$$U(x)=\frac{x}{N} \tag{5}$$

$U(x)$用来简化梯度计算，N是$\boldsymbol{X}$的时空位置数。DA-NL-1如式(6)和(7)所示：

$$\boldsymbol{Y}^{\mathrm{DA-NL-1}}=\frac{1}{N}\cos\left(\boldsymbol{Q},\boldsymbol{K}\right)\boldsymbol{V}=\frac{1}{N}\cos\left(\boldsymbol{X}\boldsymbol{W}_\theta,\boldsymbol{X}\boldsymbol{W}_\phi\right)\boldsymbol{X}\boldsymbol{W}_g \tag{6}$$

$$\boldsymbol{Z}^{\mathrm{DA-NL-1}}=\boldsymbol{X}+\boldsymbol{W}_z\boldsymbol{Y}^{\mathrm{DA-NL-1}} \tag{7}$$

式中：$\boldsymbol{W}_z\in\mathbb{R}^{C_v\times C}$将$\boldsymbol{Y}^{\mathrm{DA-NL-1}}$转换为与$\boldsymbol{X}$相同的通道维度。$\boldsymbol{Z}^{\mathrm{DA-NL-1}}$中包含了原始$\boldsymbol{X}$的特征信息和通过余弦相似性重新校准的各时空位置的特征。余弦相似性是每个时空位置贡献相似度权重，权重只包含双方位置间的$\cos\theta_{ij}$值，学习的权重从整体上也较小，能更直接地最大化相似特征。

(2) DA-NL-2：与DA-NL-1区别在于归一化函数*softmax*，如式(8)所示，*softmax*应用在$\boldsymbol{A}\odot\cos\left(\boldsymbol{Q}_i,\boldsymbol{K}_j\right)$矩阵各行中。

$$softmax\left(\boldsymbol{A}\odot\cos\left(\boldsymbol{Q}_i,\boldsymbol{K}_j\right)\right)=\frac{\mathrm{e}^{\boldsymbol{A}_{i,j}\cos\left(\boldsymbol{Q}_i,\boldsymbol{K}_j\right)}}{\sum_{n=1}^{N}\mathrm{e}^{\boldsymbol{A}_{i,n}\cos\left(\boldsymbol{Q}_i,\boldsymbol{K}_n\right)}} \tag{8}$$

DA-NL-2如式(9)(10)所示：

$$\boldsymbol{Y}^{\mathrm{DA-NL-2}}=softmax\left(\boldsymbol{A}\odot\cos\left(\boldsymbol{Q},\boldsymbol{K}\right)\right)\boldsymbol{V}=softmax\left(\boldsymbol{A}\odot\cos\left(\boldsymbol{X}\boldsymbol{W}_\theta,\boldsymbol{X}\boldsymbol{W}_\phi\right)\right)\boldsymbol{X}\boldsymbol{W}_g \tag{9}$$

$$\boldsymbol{Z}^{\mathrm{DA-NL-2}}=\boldsymbol{X}+\boldsymbol{W}_z\boldsymbol{Y}^{\mathrm{DA-NL-2}} \tag{10}$$

其中，余弦相似度运算的结果范围为−1~1。

仅使用余弦相似度项作为*softmax*函数的输入，可能不足以对条件概率分布进行建模。因此，为了增加每个条件概率项的可能概率值的范围，添加了缩放超参数矩阵$\boldsymbol{A}$，⊙表示哈玛点积乘法。

2.2 时空分解定向注意力模块

DA-NL计算复杂度为$O(N^2)$，而AR任务中$N=T\times H\times W$，并且包含去除绝对值操作的DA-NL计算量较NL更高。因此，本节对DA-NL时空分解为DA-DA(directed attention-directed attention)。

(1) 时空分解表示。前文将输入特征$\mathcal{X}\in\mathbb{R}^{T\times H\times W\times C}$折叠表示为$\boldsymbol{X}\in\mathbb{R}^{N\times C}$，本节将$\boldsymbol{X}$时空分解表示为$\boldsymbol{X}\in\mathbb{R}^{(S,T)\times C}$。$\boldsymbol{X}_{s,t}$表示$\boldsymbol{X}$在时间位置$t$和空间位置$s$的$C$维向量，其中$1\leqslant s\leqslant S(S=H\times W)$，$1\leqslant t\leqslant T$。

如图2所示，符号‘$\boldsymbol{C}^t$’和‘$\boldsymbol{C}^s$’分别表示时间和空间的余弦相似性运算，‘$\times^t$’和‘$\times^s$’表示时间和空间维度点积乘法运算。灰色矩形表示核大小为1的卷积和归一化函数，以DA-NL-1为例，基于$\boldsymbol{X}_{s,t}$将其形式化为

$$\begin{cases}\boldsymbol{Q}_{s,t}=\omega_\theta\left(\boldsymbol{X}_{s,t}\right)\\ \boldsymbol{K}_{s,t}=\omega_\phi\left(\boldsymbol{X}_{s,t}\right)\\ \boldsymbol{V}_{s,t}=\omega_g\left(\boldsymbol{X}_{s,t}\right)\end{cases} \tag{11}$$

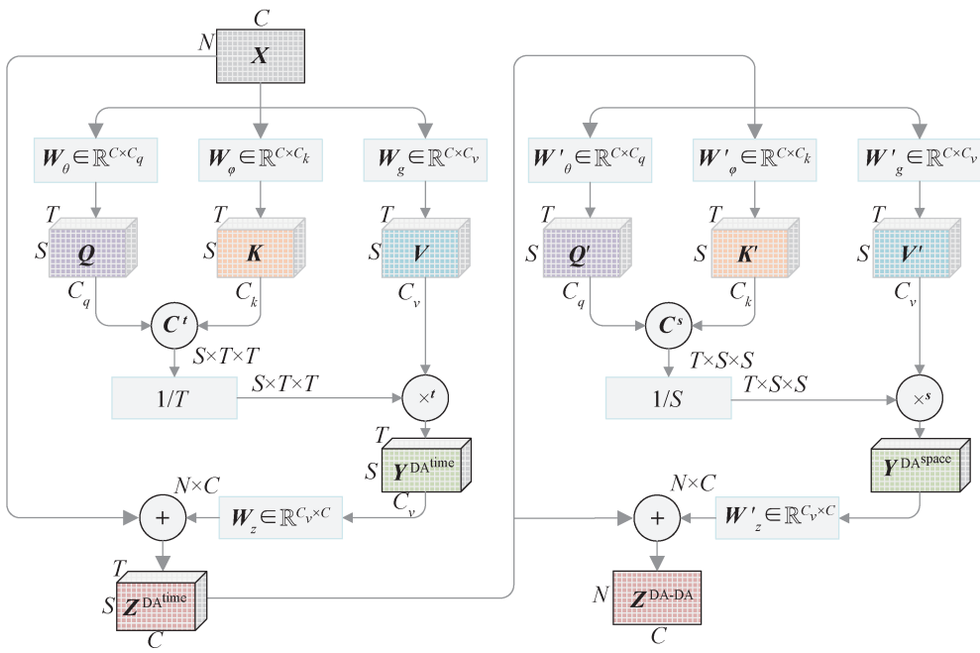


图 2 DA-DA 时空分解定向注意力非局部块

Fig. 2 Spatio-temporal decomposition of directed attention non-local moduleblock DA-DA

$$\boldsymbol{Y}_{s,t}=\sum_{s'=1}^{S}\sum_{t'=1}^{T}\frac{1}{N}\cos\left(\boldsymbol{Q}_{s,t},\boldsymbol{K}_{s',t'}\right)\boldsymbol{V}_{s',t'} \tag{12}$$

$$\boldsymbol{Z}_{s,t}=\boldsymbol{X}_{s,t}+\omega_z\left(\boldsymbol{Y}_{s,t}\right) \tag{13}$$

式中：ω_θ、ω_ϕ、ω_g、ω_z 是对应于 $\boldsymbol{W}_\theta$、$\boldsymbol{W}_\phi$、$\boldsymbol{W}_g$、$\boldsymbol{W}_z$ 的转换函数。$\boldsymbol{Y}_{s,t}$ 中需同时考虑时空维度，计算成本较大，因此将定向注意力在时间和空间维度上分解操作。

(2) $\mathrm{DA}^{\mathrm{time}}$。将各空间位置的特征独立，DA-NL 限制在时间维度中，得到各空间位置的时间定向注意力特征，形式化为

$$\boldsymbol{Y}_{s,t}^{\mathrm{time}}=\sum_{t'=1}^{T}\frac{1}{T}\cos\left(\boldsymbol{Q}_{s,t},\boldsymbol{K}_{s,t'}\right)\boldsymbol{V}_{s,t'} \tag{14}$$

$$\boldsymbol{Z}_{s,t}^{\mathrm{time}}=\boldsymbol{X}_{s,t}+\omega_z\left(\boldsymbol{Y}_{s,t}^{\mathrm{time}}\right) \tag{15}$$

(3) $\mathrm{DA}^{\mathrm{space}}$。将各时间位置的特征独立，DA-NL 限制在空间维度中，得到各时间位置的空间定向注意力特征，形式化为

$$\begin{cases}\boldsymbol{Q}'_{s,t}=\omega'_\theta\left(\boldsymbol{Z}_{s,t}^{\mathrm{time}}\right)\\ \boldsymbol{K}'_{s,t}=\omega'_\phi\left(\boldsymbol{Z}_{s,t}^{\mathrm{time}}\right)\\ \boldsymbol{V}'_{s,t}=\omega'_g\left(\boldsymbol{Z}_{s,t}^{\mathrm{time}}\right)\end{cases} \tag{16}$$

$$\boldsymbol{Y}_{s,t}^{\mathrm{space}}=\sum_{s'=1}^{S}\frac{1}{S}\cos\left(\boldsymbol{Q}'_{s,t},\boldsymbol{K}'_{s',t}\right)\boldsymbol{V}'_{s',t} \tag{17}$$

$$\boldsymbol{Z}_{s,t}=\boldsymbol{Z}_{s,t}^{\mathrm{time}}+\omega'_z\left(\boldsymbol{Y}_{s,t}^{\mathrm{space}}\right) \tag{18}$$

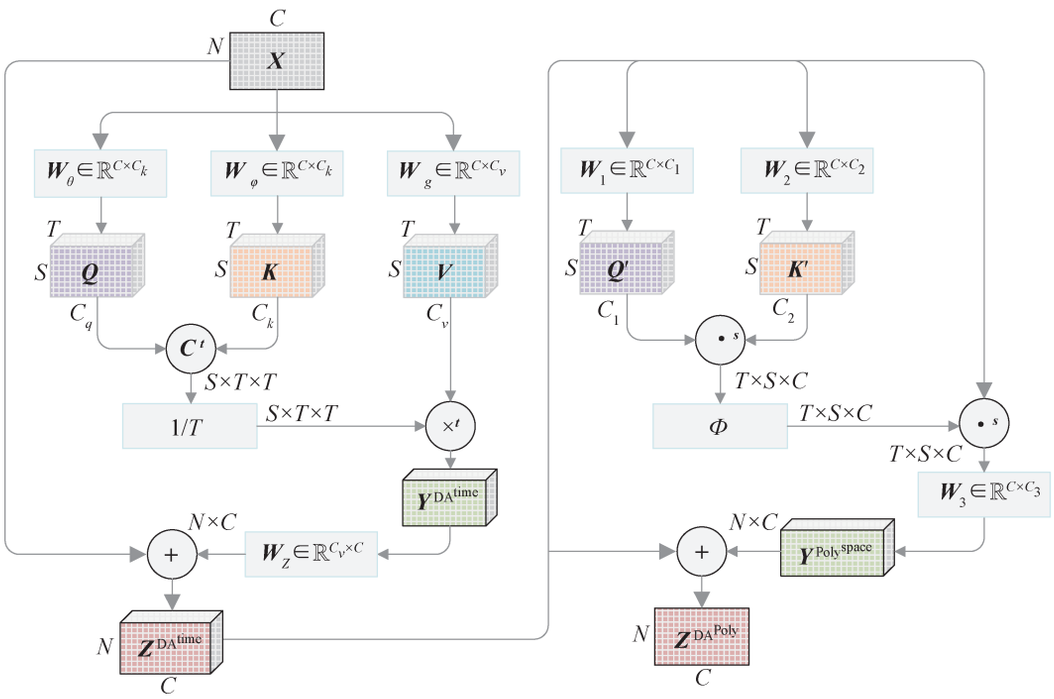
式中：ω'_θ、ω'_ϕ、ω'_g 将 $\boldsymbol{Z}_{s,t}^{\mathrm{time}}$ 映射到新的查询特征 $\boldsymbol{Q}'_{s,t}$、键值特征 $\boldsymbol{K}'_{s,t}$ 和权值特征 $\boldsymbol{V}'_{s,t}$。$\boldsymbol{Y}_{s,t}^{\mathrm{space}}$ 通过与 $\boldsymbol{Z}_{s,t}^{\mathrm{time}}$ 残差连接形成最终的时空定向注意力特征图。

2.3 高效组合定向注意力模块

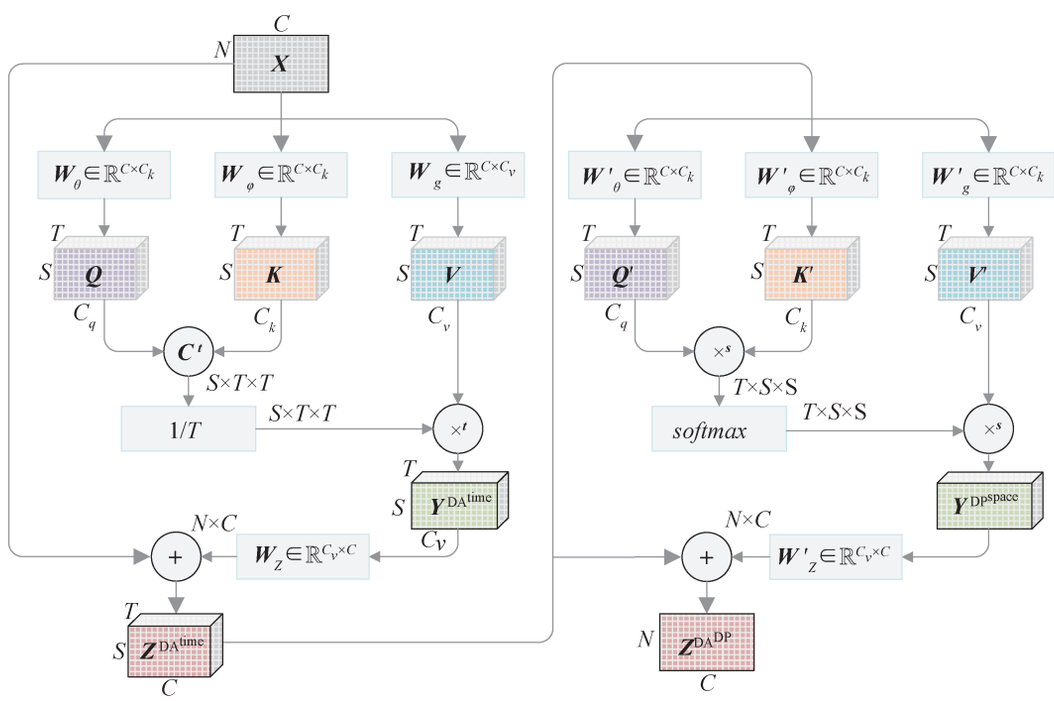
Poly-NL[3]和 NL[1]的计算量较 DA-NL 更小。因此，如图 3 所示(图 3(a)DA-Poly，对时空维度分别执行余弦相似性和哈玛点积运算点积乘法运算，‘$\boldsymbol{C}^t$’表示时间维度余弦相似性运算，‘$\bullet^s$’表示空间维度哈玛点积运算；图 3(b)DA-DP，对时空维度分别执行余弦相似性和点积乘法运算，‘$\times^s$’表示空间维度点积乘法运算)，在 DA-DA 基础上，结合 Poly-NL[3]、NL[1]共设计了 6 个定向注意力模块以改进效率，并在第 4 节进行了消融实验。

(1) DA-Poly：首先，在时间维度进行 DA-NL，如式(14)和(15)所示。

而后，在空间维度进行 Poly-NL 运算，如式(19)所示：



(a) DA- Poly



(b) DA-DP

图 3　高效组合定向注意力模块 DA-Poly 和 DA-DP

Fig. 3　Efficiently combined directed attention module DA-Poly and DA-DP

$$\boldsymbol{Y}_{s,t}^{\text{space}}=\omega_3\left(\omega_{\text{e}}^{\text{space}}\left(\sum_{s'=1}^{S}\frac{1}{S}\times\omega_1\left(\boldsymbol{Z}_{s',t}^{\text{time}}\right)\odot\omega_2\left(\boldsymbol{Z}_{s',t}^{\text{time}}\right)\right)\odot\boldsymbol{Z}_{s,t}^{\text{time}}\right) \tag{19}$$

式中：ω_1、ω_2、ω_3为分别对应于$\boldsymbol{W}_1$、$\boldsymbol{W}_2$、$\boldsymbol{W}_3$的转换函数($\boldsymbol{W}_1\in\mathbb{R}^{C\times C_1}$, $\boldsymbol{W}_2\in\mathbb{R}^{C\times C_2}$, $\boldsymbol{W}_3\in\mathbb{R}^{C\times C_3}$是 3 个可学习参数线性矩阵)；⊙表示哈玛点积乘法；$\omega_{\text{e}}^{\text{space}}$是空间扩展功能将$\mathbb{R}^{1\times T\times C}\to\mathbb{R}^{S\times T\times C}$。$\omega_3$整合

最终的各位置特征，$\boldsymbol{XW}_1$与$\boldsymbol{XW}_2$具有相同的维度即$C_1=C_2$。

(2) DA-DP：与DA-Poly类似，同样在时间维度进行DA-NL，但在空间维度进行NL，如式(16)和(20)所示：

$$\boldsymbol{Y}_{s,t}^{\text{space}}=\sum_{s'=1}^{S}\text{softmax}\left(\boldsymbol{Q}'_{s,t}\boldsymbol{K}'_{s',t}\right)\boldsymbol{V}'_{s',t} \tag{20}$$

(3) Poly-DA：首先，在时间维度进行Poly-NL，如式(21)所示：

$$\boldsymbol{Y}_{s,t}^{\text{time}}=\omega_3\left(\omega_{\text{e}}^{\text{time}}\left(\sum_{t'=1}^{T}\frac{1}{T}\times\omega_1\left(\boldsymbol{X}_{s,t'}\right)\odot\omega_2\left(\boldsymbol{X}_{s,t'}\right)\right)\odot\boldsymbol{X}_{s,t}\right) \tag{21}$$

$$\boldsymbol{Z}_{s,t}^{\text{time}}=\boldsymbol{X}_{s,t}+\omega_z\left(\boldsymbol{Y}_{s,t}^{\text{time}}\right)$$

式中：$\omega_{\text{e}}^{\text{time}}$是时间扩展功能将$\mathbb{R}^{S\times1\times C}\to\mathbb{R}^{S\times T\times C}$。

而后，在空间维度进行DA-NL，即式(16)和(17)。

(4) DP-DA：首先，在时间维度进行NL，如式(22)所示：

$$\boldsymbol{Y}_{s,t}^{\text{time}}=\sum_{t'=1}^{T}\text{softmax}\left(\boldsymbol{Q}_{s,t},\boldsymbol{K}_{s,t'}\right)\boldsymbol{V}_{s,t'} \tag{22}$$

$$\boldsymbol{Z}_{s,t}^{\text{time}}=\boldsymbol{X}_{s,t}+\omega_z\left(\boldsymbol{Y}_{s,t}^{\text{time}}\right)$$

而后，在空间维度进行DA-NL，如式(16)和(17)所示。

3 定向注意力行为识别模型

本节使用了文献[1]中使用的ResNet-C2D和ResNet-C3D基准网络，以便各版本定向注意力模块插入在CNN中形成DAM。

3.1 行为识别基准架构

(1) ResNet-C2D。2D卷积具有强大的空间特征表示能力，并能通过池化层简单聚合时间特征。因此选择2D ResNet[34]为骨干网络，并且对其调整以处理视频数据[1]。以ResNet50为例，其适用于视频行为识别的ResNet50-C2D版本如表2所示(对每个块的第一个残差单元执行空间跨步[1, 34])。ResNet50-C2D较原始ResNet-50的差异是，所有$k\times k$的2D卷积都调整为$1\times k\times k$的3D卷积形式，并在Conv_1之后添加时间池化Pool_1将帧数减半，全局平均池Pool_2从7×7调整为4×7×7使得空间和时间特征同时聚合。类似的，基于原始ResNet101改进了ResNet101-C2D。

表2 ResNet50-C2D架构
Table 2 Architecture of ResNet50-C2D

名称	输出尺寸	卷积核	步长
输入	32×224×224		
Conv_1	16×112×112	1×7×7，64	1，2，2
Pool_1	8×56×56	3×3×3，max	1，2，2
Res_1	8×56×56	$\begin{bmatrix}1\times1\times1, & 64\\ 1\times3\times3, & 64\\ 1\times1\times1, & 256\end{bmatrix}\times3$	1，1，1
Pool_2	4×56×56	3×1×1, max	2，1，1
Res_2	4×28×28	$\begin{bmatrix}1\times1\times1, & 128\\ 1\times3\times3, & 128\\ 1\times1\times1, & 512\end{bmatrix}\times4$	1，2，2
Res_3	4×14×14	$\begin{bmatrix}1\times1\times1, & 256\\ 1\times3\times3, & 256\\ 1\times1\times1, & 1024\end{bmatrix}\times6$	1，2，2
Res_4	4×7×7	$\begin{bmatrix}1\times1\times1, & 512\\ 1\times3\times3, & 512\\ 1\times1\times1, & 2048\end{bmatrix}\times3$	1,2,2
Pool_3	1×1×1	4×7×7, average	1，1，1
Fc	1×1×1	2048×class	1，1，1

(2) ResNet-C3D。通过引入额外的时间维度T，将维度$H\times W$的2D卷积转换为维度$T\times H\times W$的3D卷积。传统ResNet中有两种主要类型的卷积核，即1×1和3×3。给定一个由1×1和3×3卷积组成的残差单元，可将中间的3×3卷积核膨胀为3×3×3($\text{C3D}_{3\times3\times3}$)[10]，或将第一个1×1卷积核膨胀成3×1×1($\text{C3D}_{3\times1\times1}$)[7]。文献[1]表明，$\text{C3D}_{3\times3\times3}$和$\text{C3D}_{3\times1\times1}$实现了相当性能，而后者参数较少，计算效率更高。因此ResNet-C3D采用$\text{C3D}_{3\times1\times1}$版本。

3.2 定向注意力模型构建

基于ResNet-C2D，插入各定向注意力模块构建DAM。如图4所示，DAM是根据4.3节的消融实验设计，基于DA-NL、DA-DA和DA-Poly模块

DOI: 10.16182/j.issn1004731x.joss.22-0937

构建了多模块DAM。预处理后的视频帧输入到ResNet-C2D基础架构中对每个帧单独处理，最后对各帧的结果进行平均得到最终分类结果。在相同2D CNN框架下，DAM较C3D具有更少的参数和计算成本。在卷积层的推理过程中，特征图通过2D CNN进行独立运行，不同之处是在架构的不同位置插入了若干定向注意力模块，较3D卷积而言能够提取到更强大的时空定向关系特征。DAM的独特优势是它可轻松地将任何现成的2D CNN模型转换为可以同时处理空间和时间信息的伪3D定向全局模型，与传统的C3D相比，其在降低计算量的同时提取到了时空的定向全局关系。

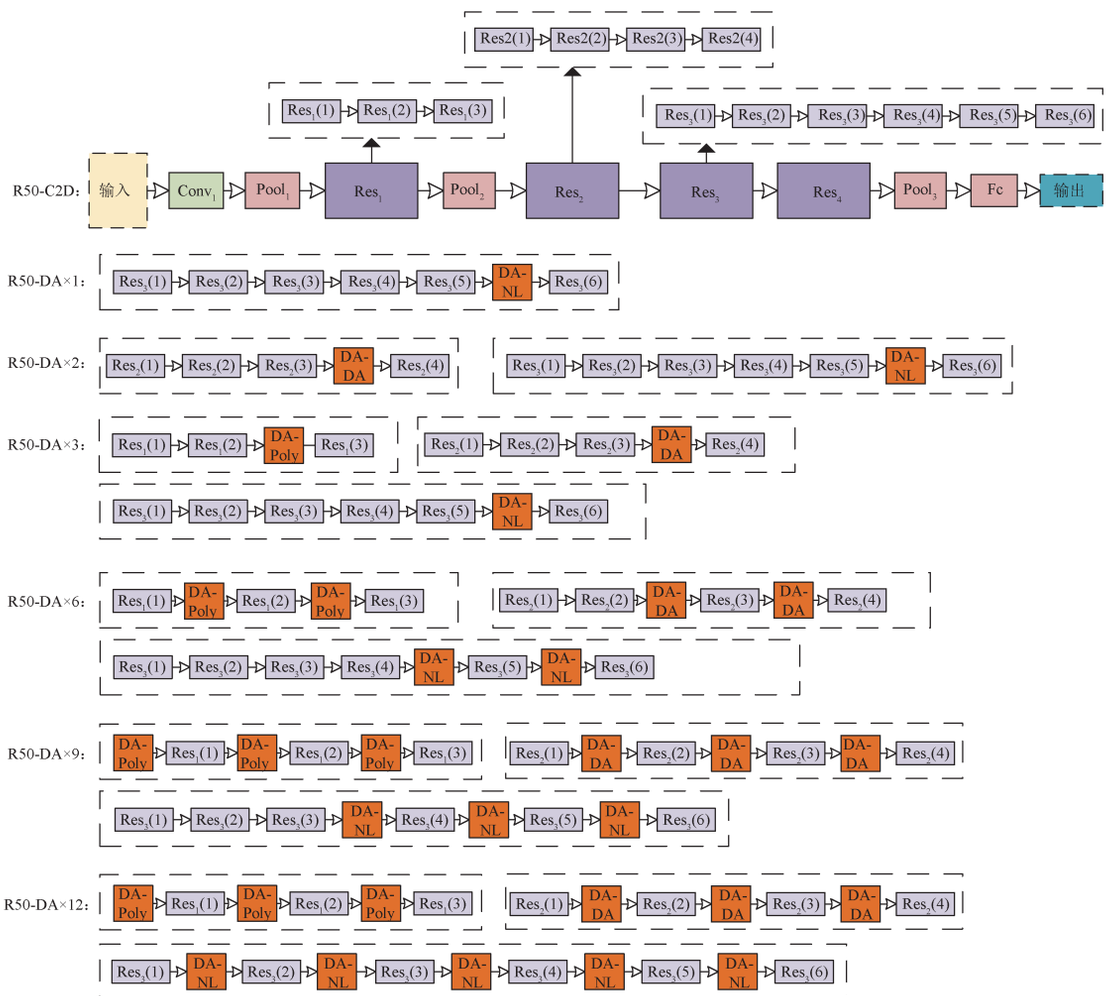


图4　基于不同数量定向模块的DAM

Fig. 4　DAM based on different number of orientation modules

4　实验评估

本节首先通过消融实验以在ResNet中合理使用各定向注意力模块，从而构造DAM；进而将DAM与流行AR模型进行比较；其次验证DAM的迁移学习能力。

4.1　数据集

Mini-Kinetics-400数据集。考虑到数据量和硬件条件，为提升实验速度基于Kinetics-400[35]轻量改造。包含400个动作类，共有23 156个训练视频和1 862个验证视频，共12 G。

Something-Something V1[36](Sth-Sth V1)数据集

带有大量动作标签，更强调动作本身，包含人对日常对象执行的基本动作，动作类有 174 个，108 499 个视频，持续时间 2~6 s。

UCF101[37]共有 13 320 个剪辑视频，101 个动作类，一个动作类的剪辑分为 25 组，每组包含 4-7 个剪辑，组中的剪辑具有一些共同的特征，例如背景或演员。动作类别包括人物交互、单人动作、人人交互、演奏乐器、运动。HMDB51[38]共包含 6 849 个剪辑视频，分为 51 个动作类别，每类别至少包含 101 个剪辑。类别包括面部动作、一般身体动作、物体交互动作、人类互动动作。本文使用 three train/test splits 策略[37]拆分 UCF101 和 HMDB51 数据集，以保持测试的一致性。此训练/测试拆分的设计方式是保持组分开，组内的剪辑是从单个长视频中获得的，因此在训练和测试中不共享来自同一组的剪辑。为了保证结果的客观准确，实验中显示了测试准确率的平均值。

4.2 设置细节

DAM 是在 Ubuntu 平台上使用 Pytorch 实现，具有一个 NVIDIA RTX 3090 显卡(24 GB RAM)。

(1) 训练。使用 SGDM(SGD with Momentum)优化器，以最小化损失并更新包括权重、偏差等在内的参数。训练时为了获得更好的性能并加快处理速度，初始学习率设置为 0.001，使用 ExponentialLR 按指数衰减调整学习率，dropout 设置为 0.5 以减缓过拟合，动量设置为 0.9，L2 正则化权重衰减为 5×10^{-4}。视频输入使用 32 个采样帧训练模型，这些采样帧是通过从原始视频中采样 64 帧，然后平均分成 32 个片段，从每个片段中挑选出一帧总共 32 帧。这些帧的较短边的大小固定为 256，并利用边角裁剪和尺度抖动来增强数据，每个裁剪后的帧最终调整为 224×224。用于训练模型的输入大小为 $N\times32\times3\times224\times224$，其中 N 是 batch 大小，N 设置为 16，所有模型都经过 100 个 epoch 的训练以进行微调，并选择最高的测试准确度进行报告。

(2) 测试。在消融实验中使用 three-crop 策略[26]进行测试，首先将每帧的较短边缩放为 256，并从缩放帧中截取 256×256 的 3 个裁剪，从全长视频中随机采样 10 次，最终预测是所有剪辑的平均 *softmax* 分数。其他实验使用单采样和单中心 224×224 中心裁剪。

4.3 消融实验

本节对定向注意力模块在 Mini-Kinetics-400 数据集上消融实验。表 3 呈现了基于不同关系函数的单模块性能。表 4 和图 5 表现了各版本模块在不同位置上的准确性和效率。图 6 展示了基于各模块在不同数量情况下的各项性能，以构造 DAM。

表 3 基于不同关系函数的单模块性能对比
Table 3 Single module performance comparison based on different relationship functions

模块	参数量/MB	GFLOPS×View	Top1/%	Top5/%
基准模型	24.18	26.19×10×3	71.76	89.85
DP-NL[1]	25.56	26.47×10×3	72.68	90.47
Poly-NL[3]	25.33	26.20×10×3	72.65	90.46
DA-NL-1	25.53	27.02×10×3	73.21	90.73
DA-NL-2	25.57	27.04×10×3	73.18	90.71

如表 3 所示(插入位置是 Res_3 中最后一个单元的前面，架构：Res50-C2D，帧数×view：32×256×256×10×3)，DA-NL 在识别精度上较 DP-NL[1]和 Poly-NL[3]有明显提升，说明基于余弦相似性的定向注意力可更为有效表征各时空位置特征间的关系。但 DA-NL 较 DP-NL 和 Poly-NL 计算量稍大，不利于通过多帧输入和多模块途径，进一步提高识别精度。另外，DA-NL-1 和 DA-NL-2 在精确度上未存在明显差异，这也许是因定向注意力去除了绝对值，不需 *softmax* 函数对向量幅度弱化。此后，DA-NL 均为 DA-NL-1。

表 4(插入位置是各 Res 块中最后一个单元的前面，输入：32×256×256×10×3，架构：Res50-C2D)比较了不同 Res 层的各定向注意力时空分解模块：DA-NL、DA-DA、DA-Poly、DA-DP、Poly-DA、DP-DA。

DOI: 10.16182/j.issn1004731x.joss.22-0937

表 4 各组合定向注意力模块在Res层的性能比较

Table 4 Performance comparison of each combined directed attention module in Res layer

位置	模块	参数量/MB	GFLOPS×View	Top1/%	Top5/%
	基准模型	24.18	26.19×10×3	71.76	89.85
Res_1	DA-NL	24.27	**237.92**×10×3	**72.94**	**90.43**
	DA-DA	25.19	52.72×10×3	72.72	90.35
	DA-Poly	24.83	**26.26**×10×3	**72.43**	90.27
	DA-DP	25.22	35.08×10×3	72.46	**90.26**
	Poly-DA	24.83	52.66×10×3	72.68	90.32
	DP-DA	25.22	52.68×10×3	72.65	90.31
Res_2	DA-NL	24.51	**32.81**×10×3	**73.07**	**90.58**
	DA-DA	26.23	27.85×10×3	72.94	90.41
	DA-Poly	26.08	**26.20**×10×3	72.68	90.32
	DA-DP	26.25	26.75×10×3	**72.65**	**90.31**
	Poly-DA	26.08	27.84×10×3	72.90	90.38
	DP-DA	26.25	27.84×10×3	72.91	90.37
Res_3	DA-NL	25.53	**27.02**×10×3	**73.21**	**90.73**
	DA-DA	27.98	26.40×10×3	72.84	90.53
	DA-Poly	27.35	**26.19**×10×3	72.53	90.39
	DA-DP	30.02	26.26×10×3	**72.49**	**90.37**
	Poly-DA	27.35	26.39×10×3	72.76	90.40
	DP-DA	30.02	26.39×10×3	72.74	90.39
Res_4	DA-NL	29.68	**26.29**×10×3	**72.51**	**90.25**
	DA-DA	36.18	26.22×10×3	72.34	90.15
	DA-Poly	35.82	**26.19**×10×3	**72.01**	**89.98**
	DA-DP	36.24	26.20×10×3	72.04	90.02
	Poly-DA	35.82	26.22×10×3	72.28	90.12
	DP-DA	36.24	26.22×10×3	72.31	90.10

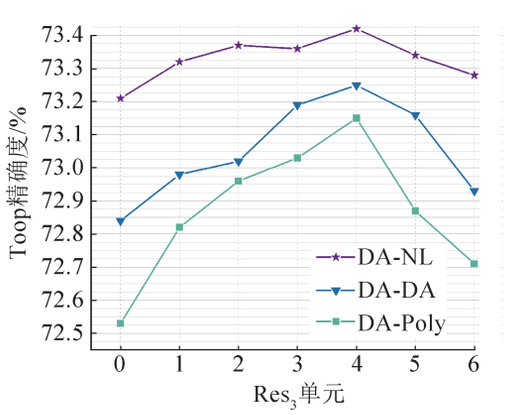


图 5 在Res_3插入双模块的不同方案

Fig. 5 Different options for inserting dual modules in Res_3

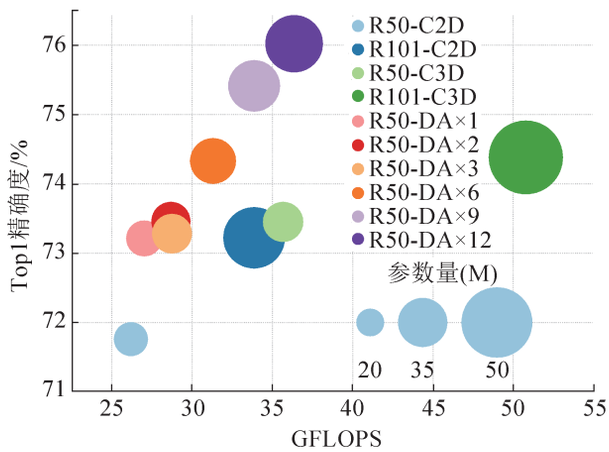


图 6 DAM与基准架构的比较

Fig. 6 Comparison of DAM with the benchmark architectures

根据Res层纵向分析，ResNet50-C2D的性能指标有以下规律。参数量随模块插入位置的深度增加而逐渐提升，这可能是因深层特征通道数C较浅层特征变大，导致模块中变换矩阵参数量更大；计算量随插入层位置深度增加而逐渐降低，这可能是深层特征时空位置N逐渐减小的原因；精确度随插入位置深度增加逐渐提升，并在Res_4层出现了回落，这可能是因Res_4特征时空位置太少，导致特征间的关系特征提取不充分，不能充分发挥注意力机制效能。

另外，各Res层中出现共性情况。DA-NL精确度最高，说明其提取特征间注意力关系最为充分；DA-DA较DA-NL的计算量明显降低，精确度仅次于未分解的DA-NL，证明时空分解操作的有效性；DA-Poly的计算量最低且保证了精度，说明定向注意力和线性注意力结合的高效性；Poly-DA和DP-DA较DA-DA，并未在效率和精度上表现明显优势，这可能是因特征图的空间位置远大于时间位置，所以仅在时间维度进行Poly-NL不足以改进计算量。

其中：在Res_1层，DA-NL较基准架构精度提升了1.22%，但GFLOPS高达237.92即ResNet50-C2D的近10倍计算量，这显然不能满足实际应用；DA-DA将GFLOPS降低到了基准线的2倍即52.72，并具备1.02%的精度提升；DA-Poly的

GFLOPS 与基线几乎相同即 26.26，同时具有 0.67% 精度提升；在 Res_2 层，DA-DA 和 DA-Poly 较 DA-NL 在计算量上分别降低了 4.96 和 6.61，在精确度上分别降低了 0.13 和 0.42。Res_3 和 Res_4 层，DA-DA 和 DA-Poly 较 DA-NL 计算量降低并不明显。

图 5(横坐标‘0’：模块 1 插入在第 6 个 Res_3 单元前，横坐标‘1~6’：第 2 个模块插入在 Res_3 单元前的位置，输入：32×256×256×30，架构：Res50-C2D)呈现了 Res_3 中插入双模块的不同方案。无间隔插入(Res_3 单元为 6)的精确度不如间隔插入(Res_3 单元为 1~5)的精度高，其中间隔 1 个 res 块(Res_3 单元为 5)和间隔 2 个 res 块(Res_3 单元为 4)的增益较好。这可能是因无间隔插入缺少了卷积的归纳偏置能力，导致双模块提取的关系特征较单模块增益不明显。

图 6 呈现了模块数量为 1、3、6、9、12 情况下的 DAM 效能。DAM 的精确度随着模块的使用数量逐步提升，其中：DA×1 的 Top1 精确度与 R101-C2D 相当，且参数量和 GFLOPS 更低，证明了使用单个定向注意力模块比加深 CNN 的深度更有利于提升识别效果，并且效率更高。DA×6 的精确度超过 R50-C3D，计算量是 R50-C3D 的 87%，参数量是其 113%，说明较于传统 3D 卷积，使用定向注意力模块能够取得更高的精确度以及计算效率。DA×9 的精确度高于 R101-C3D，且计算量是其 67%，参数量是其 70%。综上所述，多定向注意力模块设计的 DAM，较传统的网络延伸和 3D 卷积方法，在精度和效率上都具备优势。

4.4 与现有技术的比较

为了与其他方法公平比较，在 Sth-Sth V1 数据集[36]上使用单帧和单中心 224×224 裁剪的测试方案，预训练使用 ImageNet 数据集[39]和 Kinetics-400 数据集[35]。DAM 采用了图 6 中 12 模块的版本，并测试了 8、16 和 32 帧的 3 种情况选择的对比模型算法，均是使用了 ResNet50 作为基准架构，具备相同的网络深度，计算量和参数量都在同一层级，便于比较各模型性能。另外，选择的对比模型，涉及 AR 领域的多个设计方向，包括：纯 2D 卷积的 C2D、具备光流的 TSN；基于 3D 卷积的 I3D、时空分解卷积的 S3D 和混合卷积的 ECO；基于时态模块的 TRN、TSM、TEA；基于自注意机制的 TDN 和 NL。各对比模型算法的实验训练参数依照原配置，如表 5 所示(“*”表示未发布源代码，“I”表示 ImagNet，“K400”表示 Kinetics-400，“S1M”表示 Sports1M)。

如表 6 所示，传统 2D CNN 方法中，C2D[1]的 Top1 精确度仅为 18.7%，TSN[13]虽利用稀疏采样策略提取全局特征，但精确度仅为 19.7%，可知 2D 卷积时间建模较弱。而 DAM 基于定向注意力机制，在 2D CNN 架构下实现了 48.2%~51.7% 的精确度，较 C2D 和 TSN 实现了极大提升，说明定向注意力较 2D 卷积具备较强的时间特征提取能力。另外，3D CNN 方法中，C3D[1]、ECO[17]、I3D[40]和 S3D-G[15]虽然较传统 2D CNN 具备一定的时间特征提取能力，但其精确度较 DAM 仍有一定差距，其中只有 S3D-G[15]在识别效果上与 DAM 相近，但其计算量远大于 8 帧输入的 DAM，说明 3D 卷积只是实现了局部时空特征融合，而 DAM 通过定向注意力更好地实现了全局时空特征融合。最后在时态卷积方法中，在同样的 32 帧输入下，DAM 的精确度较 I3D+NL[40]高出 7.3%，而计算量仅需其 10%，说明定向注意力较传统注意力具备更好的全局特征提取能力，以及各模块的搭配设计使得效率得到了很好地优化。DAM 较 TRN[12]和 TSM[11]识别效果更优异，较 TEA[21]和 TDN[26]在计算量上更为轻量，但精确度还有一定差距。

4.5 UCF101 的迁移学习

本节将 4.4 节预训练的 DAM 转移到更小的 UCF101[37]和 HMDB51[38]数据集，并在 UCF101 和 HMDB51 上进行了微调，以表明 DAM 学习的表示可推广到其他数据集。表 7 中包含的结果是 three train/test splits 平均类准确度[37]。很明显，DAM 展示了非常强大的迁移学习能力，平均分类准确率高达 95.7% 和 71.8%。

DOI: 10.16182/j.issn1004731x.joss.22-0937

表 5 各对比模型在 Sth-Sth V1 数据集上的实验训练参数

Table 5 Experimental training parameters of each comparison model on Sth-Sth V1 dataset

模型	主干架构	预训练	初始学习率	Dropout	动量	权重衰减	帧数	分辨率	周期	Batch
C2D[1]	ResNet50	I+K400	0.001	0.5	0.9	5×10^{-4}	32	224×224	100	16
TSN[13]	BNInception	I	0.020	0.8	0.9	1×10^{-4}	8	224×224	50	8
TSN[13]	ResNet50	I	0.020	0.8	0.9	1×10^{-4}	8	224×224	50	8
C3D[1]	ResNet50	I+K400	0.001	0.5	0.9	5×10^{-4}	32	224×224	100	16
ECO[17]	BNIncep+R18	I+K400	0.001	0.5	0.9	5×10^{-4}	8	224×224	100	32
I3D[40]	ResNet50	I+K400	0.001 25	0.3	0.9	1×10^{-4}	32	224×224	100	8
I3D+NL[40]	ResNet50	I+K400	0.001 25	0.3	0.9	1×10^{-4}	32	224×224	100	8
S3D-G[15]	BN-Inception	I	0.100	0.5	0.9	1×10^{-4}	64	224×224	100	6
TRN[12]	BNInception	I	0.001	0.5	0.9	1×10^{-4}	8	224×224	100	10
TRN[12]	ResNet50	I	0.001	0.5	0.9	1×10^{-4}	8	224×224	100	10
TSM[11]	ResNet50	I+K400	0.010	0.5	0.9	1×10^{-4}	8	224×224	50	64
TEA*[21]	ResNet50	I					16	224×224		
TDN[26]	ResNet50	I+S1M	0.020	0.5	0.9	1×10^{-4}	8	224×224	60	128
DAM	ResNet50	I+K400	0.001	0.5	0.9	5×10^{-4}	32	224×224	100	16

表 6 DAM 在 Sth-Sth V1 数据集上与其他方法的测试比较

Table 6 Comparison of DAM on Sth-Sth V1 dataset with tests of other methods

模型	主干架构	帧数×View	分辨率	参数量/MB	GFLOPS×View	Top1	Top5
C2D[1]	ResNet50	32×1×1	224×224	24.18	26.19×1×1	18.7	45.3
TSN[13]	BNInception	8×1×1	224×224	10.70	16×1×1	19.5	
TSN[13]	ResNet50	8×1×1	224×224	24.30	33×1×1	19.7	46.6
C3D[1]	ResNet50	32×1×1	224×224	28.51	35.67×1×1	32.8	60.3
ECO[17]	BNIncep+R18	8×1×1	224×224	47.50	32×1×1	39.6	
I3D[40]	ResNet50	32×2×1	224×224	28.00	153×2×1	41.6	72.2
I3D+NL[40]	ResNet50	32×2×1	224×224	35.30	168×2×1	44.4	76.0
S3D-G[15]	BN-Inception	64×1×1	224×224	11.56	71.38×1×1	48.2	78.7
TRN[12]	BNInception	8×1×1	224×224	18.30	16×1×1	34.4	
TRN[12]	ResNet50	8×1×1	224×224	31.80	33×1×1	38.9	68.1
TSM[11]	ResNet50	8×1×1	224×224	24.30	33×1×1	45.6	74.2
TEA[21]	ResNet50	16×1×1	224×224		70×1×1	**51.9**	**80.3**
TDN[26]	ResNet50	8×1×1	224×224		36×1×1	**52.3**	**80.6**
DAM	ResNet50	8×1×1	224×224	40.38	9.09×1×1	48.2	76.5
DAM	ResNet50	16×1×1	224×224	40.38	18.18×1×1	49.5	78.3
DAM	ResNet50	32×1×1	224×224	40.38	36.36×1×1	**51.7**	**80.1**

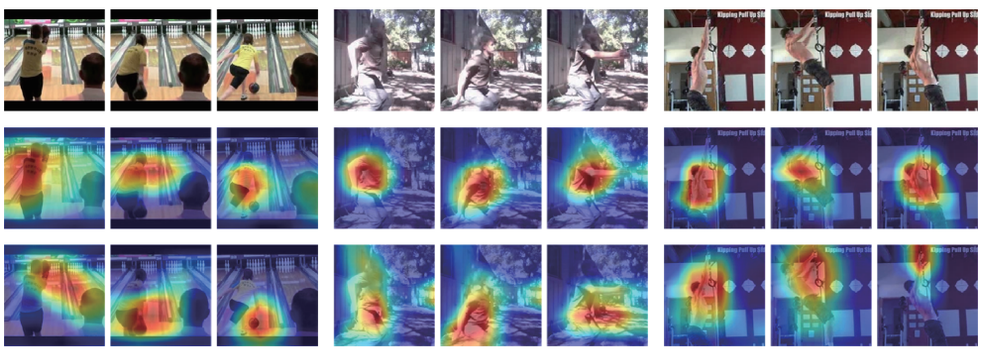
表 7 DAM 在 UCF101 和 HMDB51 上的平均类精度

Table 7 Average class accuracy of DAM on UCF101 and HMDB51

模型	预训练	UCF101	HMDB51
C3D[1]	Kinetics400	84.6	54.3
I3D[40]	Kinetics400	91.8	66.4

为了更好地理解 DAM 如何学习用于行为识别的判别性时空描述符，使用 CAM 方法[41]可视化模型的特定类激活图。在实验中，将 T 设置为 3 从 UCF101 测试集中的视频序列中均匀采样 3 个片段，以获得特定类的激活图。作为比较还可视化了 I3D 模型的激活图，该模型利用局部时空信息，而不是联合建模局部和全局时空信息。作为说明，图 7(上：原始视频帧，中：I3D，下：DAM)列出了模型的 3 个行为类别的激活图。DAM 关注人的

运动或者手部交互行为，如在打保龄球这个行为中，其重点关注了保龄球与抓球的手部。与I3D模型相比，DAM能够更精确地关注行为中的与运动位移相关的区域，而I3D缺乏对运动相关区域的精确关注，这依赖于DAM强大的时间建模能力。



(a) 打保龄球 (b) 蹲起 (c) 引体向上

图 7 使用CAM可视化特定于行为的激活图

Fig. 7 Activation diagrams of specific actions visualized by CAM

激活图表明，与无法联合建模视频中局部和全局时空特征的I3D相比，DAM可以捕获视频帧中的特征关系，其更侧重于关注与行为密切相关的时空区域。例如，DAM会更加关注附近有保龄球的人体，而只有人体的区域不会被激活。另外在“引体向上”示例中，DAM通过单杠和手部激活，然而I3D被大面积人体区域激活。

5 结论

本文提出并实现了定向注意力模型(DAM)进行AR任务，其中基于余弦相似性的定向注意力模块可去除特征绝对值从而更有效表示特征间关系，且时空分解降低了注意力运算的计算量。DAM较传统2D、3D和时态卷积，在不加大计算量的情况下表现出强大的时空特征表示能力。同时也存在一些不足：本文所提出的创新模块泛化性能和对于各模块的效率-精度平衡问题还有待进一步探索。在未来的工作中，将在更多的骨干网络上进行实验，验证所提出的模块的泛化能力；同时，将基于效率–精度平衡问题，探索如何自动生成各模块在骨干网络上的使用方案。

参考文献：

[1] Wang Xiaolong, Girshick R, Gupta A, et al. Non-local Neural Networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2018: 7794-7803.

[2] Truong T D, Quoc-Huy Bui, Chi Nhan Duong, et al. DirecFormer: A Directed Attention in Transformer Approach to Robust Action Recognition[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2022: 19998-20008.

[3] Babiloni F, Marras I, Kokkinos F, et al. Poly-NL: Linear Complexity Non-local Layers with 3rd Order Polynomials[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE, 2021: 10498-10508.

[4] Wang Heng, Cordelia Schmid. Action Recognition with Improved Trajectories[C]//2013 IEEE International Conference on Computer Vision. Piscataway, NJ, USA: IEEE, 2013: 3551-3558.

[5] Simonyan K, Zisserman A. Two-stream Convolutional Networks for Action Recognition in Videos[C]// Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge, MA, USA: MIT Press, 2014: 568-576.

[6] Christoph Feichtenhofer, Axel Pinz, Zisserman A. Convolutional Two-stream Network Fusion for Video Action Recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2016: 1933-1941.

[7] Christoph Feichtenhofer, Axel Pinz, Richard P Wildes. Spatiotemporal Residual Networks for Video Action Recognition[C]//Proceedings of the 30th International

DOI: 10.16182/j.issn1004731x.joss.22-0937

Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2016: 3476-3484.

[8] Christoph Feichtenhofer, Axel Pinz, Richard P Wildes. Spatiotemporal Multiplier Networks for Video Action Recognition[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2017: 7445-7454.

[9] Tran D, Bourdev L, Fergus R, et al. Learning Spatiotemporal Features with 3D Convolutional Networks[C]//2015 IEEE International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE, 2015: 4489-4497.

[10] João Carreira, Zisserman A. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2017: 4724-4733.

[11] Lin Ji, Gan Chuang, Han Song. TSM: Temporal Shift Module for Efficient Video Understanding[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE, 2019: 7082-7092.

[12] Zhou Bolei, Andonian A, Oliva A, et al. Temporal Relational Reasoning in Videos[C]//Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 831-846.

[13] Wang Limin, Xiong Yuanjun, Wang Zhe, et al. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition[C]//Computer Vision-ECCV 2016. Cham: Springer International Publishing, 2016: 20-36.

[14] Qiu Zhaofan, Yao Ting, Mei Tao. Learning Spatio-temporal Representation with Pseudo-3D Residual Networks[C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE, 2017: 5534-5542.

[15] Xie Saining, Sun Chen, Huang J, et al. Rethinking Spatiotemporal Feature Learning: Speed-accuracy Trade-offs in Video Classification[C]//Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 318-335.

[16] Tran D, Wang Heng, Feiszli M, et al. Video Classification with Channel-separated Convolutional Networks[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE, 2019: 5551-5560.

[17] Mohammadreza Zolfaghari, Kamaljeet Singh, Thomas Brox. ECO: Efficient Convolutional Network for Online Video Understanding[C]//Computer Vision-ECCV 2018. Cham: Springer International Publishing, 2018: 713-730.

[18] Feichtenhofer C. X3D: Expanding Architectures for Efficient Video Recognition[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2020: 200-210.

[19] Wang Xianyuan, Miao Zhenjiang, Zhang Ruyi, et al. I3D-LSTM: A New Model for Human Action Recognition [C]//IOP Conference Series: Materials Science and Engineering. Bristol, United Kingdom: IOP Publishing, 2019: 032035.

[20] Tran D, Ray J, Shou Z, et al. Convnet Architecture Search for Spatiotemporal Feature Learning[J]. Computing Research Repository, 2017, 16(8): 1-12.

[21] Li Yan, Ji Bin, Shi Xintian, et al. TEA: Temporal Excitation and Aggregation for Action Recognition[C]// 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2020: 906-915.

[22] 罗会兰, 陈翰. 时空卷积注意力网络用于动作识别[J]. 计算机工程与应用, 2023, 59(9): 150-158.
Luo Huilan, Chen Han. Spatial-temporal Convolutional Attention Network for Action Recognition[J]. Computer Engineering and Applications, 2023, 59(9): 150-158.

[23] 吴丽君, 李斌斌, 陈志聪, 等. 3D多重注意力机制下的行为识别[J]. 福州大学学报(自然科学版), 2022, 50(1): 47-53.
Wu Lijun, Li Binbin, Chen Zhicong, et al. Action Recognition Under 3D Multiple Attention Mechanism[J]. Journal of Fuzhou University(Natural Science Edition), 2022, 50(1): 47-53.

[24] Zhu Yi, Lan Zhenzhong, Newsam S, et al. Hidden Two-stream Convolutional Networks for Action Recognition [C]//Computer Vision-ACCV 2018. Cham: Springer International Publishing, 2019: 363-378.

[25] Feichtenhofer C, Fan Haoqi, Malik J, et al. SlowFast Networks for Video Recognition[C]//2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE, 2019: 6201-6210.

[26] Wang Limin, Tong Zhan, Ji Bin, et al. TDN: Temporal Difference Networks for Efficient Action Recognition [C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2021: 1895-1904.

[27] Ma C Y, Chen M H, Kira Z, et al. TS-LSTM and Temporal-inception: Exploiting Spatiotemporal Dynamics for Activity Recognition[J]. Signal Processing: Image Communication, 2019, 71: 76-87.

[28] Neimark D, Bar O, Zohar M, et al. Video Transformer Network[C]//2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Piscataway,

NJ, USA: IEEE, 2021: 3156-3165.

[29] Li Kunchang, Wang Yali, Gao Peng, et al. UniFormer: Unified Transformer for Efficient Spatiotemporal Representation Learning[EB/OL]. (2022-02-08) [2022-04-04]. https://arxiv.org/abs/2201.04676.

[30] Arnab A, Dehghani M, Heigold G, et al. ViViT: A Video Vision Transformer[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE, 2021: 6816-6826.

[31] Fan Haoqi, Xiong Bo, Mangalam K, et al. Multiscale Vision Transformers[C]//2021 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE, 2021: 6804-6815.

[32] Li Yanghao, Wu Chaoyuan, Fan Haoqi, et al. MViTv2: Improved Multiscale Vision Transformers for Classification and Detection[EB/OL]. (2022-03-30) [2022-04-04]. https://arxiv.org/abs/2112.01526.

[33] Alfasly S, Lu Jian, Xu Chen, et al. Learnable Irrelevant Modality Dropout for Multimodal Action Recognition on Modality-specific Annotated Videos[EB/OL]. (2022-03-27) [2022-04-04]. https://arxiv.org/abs/2203.03014.

[34] He Kaiming, Zhang Xiangyu, Ren Shaoqing, et al. Deep Residual Learning for Image Recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2016: 770-778.

[35] Kay W, João Carreira, Simonyan K, et al. The Kinetics Human Action Video Dataset[EB/OL]. (2017-05-19) [2022-04-04]. https://arxiv.org/abs/1705.06950.

[36] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, et al. The "Something Something" Video Database for Learning and Evaluating Visual Common Sense[C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway, NJ, USA: IEEE, 2017: 5843-5851.

[37] Soomro K, Zamir A R, Shah M. UCF101: A Dataset of 101 Human Actions Classes from Videos in the Wild[EB/OL]. (2013-12-03) [2022-04-04]. https://arxiv. org/abs/1212.0402.

[38] Kuehne H, Jhuang H, Garrote E, et al. HMDB: A Large Video Database for Human Motion Recognition[C]//2011 International Conference on Computer Vision. Piscataway, NJ, USA: IEEE, 2011: 2556-2563.

[39] Deng Jia, Dong Wei, Socher R, et al. ImageNet: A Large-scale Hierarchical Image Database[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ, USA: IEEE, 2009: 248-255.

[40] Wang Xiaolong, Gupta A. Videos as Space-time Region Graphs[C]//Computer Vision-ECCV 2018: 15th European Conference. Heidelberg: Springer-Verlag, 2018: 413-431.

[41] Zhou Bolei, Khosla A, Lapedriza A, et al. Learning Deep Features for Discriminative Localization[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ, USA: IEEE, 2016: 2921-2929.

DOI: 10.16182/j.issn1004731x.joss.22-0937