(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 116385945 B
(45)授权公告日 2023.08.25

(21)申请号 202310657865.4
(22)申请日 2023.06.06
(65)同一申请的已公布的文献号
申请公布号 CN 116385945 A
(43)申请公布日 2023.07.04
(73)专利权人 山东省人工智能研究院
地址 250000 山东省济南市历下区科院路19号
专利权人 天津理工大学
山东中联视听信息科技股份有限公司
(72)发明人 高文杰 高赞 周冕 赵一博 卓涛 李志慧 程志勇 李传森 刘冬冬
(74)专利代理机构 山东知圣律师事务所 37262
专利代理师 陈晓辉
(51)Int.Cl.
G06V 20/40(2022.01)
G06V 10/764(2022.01)
G06V 10/774(2022.01)
G06V 10/80(2022.01)
G06V 10/82(2022.01)
G06N 3/0442(2023.01)
G06N 3/0464(2023.01)
G06N 3/08(2023.01)

(56)对比文件
CN 109784269 A,2019.05.21
CN 111259795 A,2020.06.09
CN 114973416 A,2022.08.30
WO 2021051545 A1,2021.03.25
US 2022296334 A1,2022.09.22
刘文龙等.“基于多特征融合及Transformer 的人体跌倒动作检测算法”.《应用科技》.2022,第49-62页.

审查员 张思洋

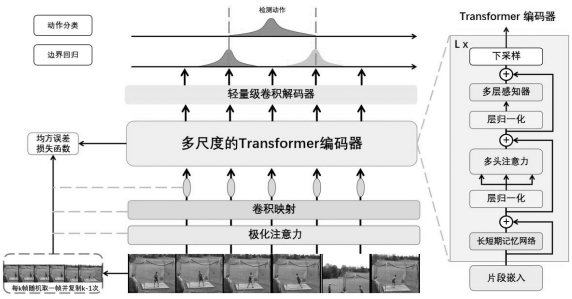
权利要求书4页 说明书9页 附图1页

(54)发明名称

基于随机帧补帧和注意力的视频交互动作检测方法及系统

(57)摘要

本发明属于计算机视觉和模式识别技术领域,尤其涉及一种基于随机帧补帧和注意力的视频交互动作检测方法及系统,方法的具体步骤如下:(1)特征提取网络的选择;(2)自注意力全局信息建模;(3)随机帧补帧数据增强;(4)金字塔特征的生成;(5)边界定位与分类。本发明能够同时聚合全局时序与多尺度的局部时序信息,通过产生的金字塔特征进行高效的动作定位。应用基于随机帧补帧进行数据增强,通过LSTM+Transformer的结合来解决单个模型在不同大小数据集上性能表现不同的问题,以获得更精确的动作定位与分类结果。



CN 116385945 B

1.一种基于随机帧补帧和注意力的视频交互动作检测方法,其特征是:包括以下步骤:

步骤10.特征提取网络的选择

选择基于Kinetics数据集预训练的I3D网络来进行特征的提取;

步骤20.自注意力全局信息建模

在步骤10基础网络选择的基础上,对全局的时序信息进行建模,对于I3D网络的输出;用Polarized Self-Attention极化注意力来寻找帧与帧之间的关系并进行加权;

在Transformer网络之前加入1D卷;

步骤30. 随机帧补帧数据增强

在第10步特征网络的输出上,将一个视频分为若干个片段,每个片段中随机取一帧,其它帧与取的帧一样,来形成一个变化较大的新特征向量;

把经过backbone的新特征向量与原视频特征向量计算一个mse损失;

步骤30中公式如下:

原视频特征向量:$X = [x_1, x_2 \dots x_T] \in R^{T*D}$,T表示视频特征序列长度,D 表示特征维度;

把X分成t/k段:$X_{seg} = \left[\overline{x}_1, \overline{x}_2 \dots \overline{x}_{\frac{T}{k}}\right]$,每个 $\overline{x}_i$ 包含k帧,i 表示第i个视频特征;

从每个片段中随机取一帧,并复制k次,

$$X_{out} = \left[\varphi(\sigma(\overline{x}_1)), \varphi(\sigma(\overline{x}_2)) \dots \varphi(\sigma(\overline{x}_{\frac{T}{k}}))\right] \in R^{T*D}$$,σ代表随机取帧,φ代表复制k次操作;

$Loss = MSE(A_1, A_2)\theta$,A_1, A_2代表向量X和X_{out}经过backbone网络之后的新的特征向量,MSE均方损失函数,θ 表示调节系数,通常为1;

步骤40.金字塔特征的生成

在步骤20步网络的基础上,将通过多尺度信息聚合模块之后的特征通过多尺度的Transformer编码成6层的特征金字塔,并且将LSTM与Transformer进行结合;

步骤50.边界定位与分类

在得到6个尺度的金字塔特征之后;对每一个尺度的金字塔特征,分别输入到不同的1D卷积中来获得定位和分类的特征,之后采用分类特征来进行分类,采用定位特征进行边界的回归,在训练分类的过程中采用focal loss进行约束,在训练回归的过程中采用GIoU loss进行约束。

2.根据权利要求1所述的基于随机帧补帧和注意力的视频交互动作检测方法,其特征是:在对于提取出来的特征通过Polarized Self-Attention中的Channel-only branch和Spatial-only branch进行操作,Channel-only branch定义如下:

$$A^{ch}(X) = F_{SG}\left[W_{z|\theta_1}\left(\sigma_1(W_v(X)) \times F_{SM}(\sigma_2(W_q(X)))\right)\right], A^{ch}(X) \in R^{C\times 1\times 1}$$,其

中W_q、W_z、W_v是1×1卷积层，σ_1，σ_2是reshape operator即把特征维度由C/2×H×W改为C/2×HW,C,表示为通道维度 ,H表示图片的高度，W表示图片的宽度，F_{SM}是softmax算子，θ_1表示通道卷积的中间参数,X是矩阵点积运算

$F_{SM}(X) = \sum_{j=1}^{N_p} \frac{e^{x_j}}{\sum_{m=1}^{N_p} e^{x_m}} x_j$,W_v、W_q和W_z之间的内部通道数是C/2,通道分支的输出是$Z^{ch} = A^{ch}(X) \odot^{ch} X \in R^{C\times H\times W}$,其中$\odot^{ch}$是通道乘法运算操作符；

Spatial-only branch定义如下：

$A^{sp}(X) = F_{SG}\left[\sigma_3\left(F_{SM}(\sigma_1(F_{GP}(W_q(X)))) \times \sigma_2(W_v(X))\right)\right]$，$A^{sp}(X) \in R^{1\times H\times W}$，其中$F_{SG}$表示Sigmoid函数$W_q$、$W_v$，是标准的1×1卷积，$\sigma_1\sigma_2\sigma_3$是三个reshape operator，$F_{SM}$是softmax算子，$F_{GP}$是全局池化操作符，

$F_{GP}(X) = \frac{1}{H\times W}\sum_{H}^{i=1}\sum_{w}^{j=1} X(:,i,j)$,空间分支的输出是

$Z^{sp} = A^{sp}(X) \odot^{sp} X \in R^{C\times H\times W}$,其中$\odot^{sp}$是空间乘法运算操作符；

通道分支和空间分支的输出在并行布局下组成：

$PSA_p(X) = Z^{ch} + Z^{sp} = A^{ch}(X) \odot^{ch} X + A^{sp}(X) \odot^{sp} X$。

3.根据权利要求1所述的基于随机帧补帧和注意力的视频交互动作检测方法,其特征是:每一个视频损失定义如下：$L = \sum_t \left(\frac{1}{T} L_{cls} + \frac{\lambda_{reg}}{T_+} 1_{ct} L_{reg}\right)$；

其中T是输入序列的长度,1_{ct}是一个指示函数,表示时间步长t是否在动作范围内,即正样本，T_+是阳性样本总数，L应用于输出金字塔上的所有级别,并在训练期间对所有视频样本进行平均，λ_{reg}是一个平衡分类损失和回归损失的系数，L_{reg}用于距离回归的一个GIoU loss,Lcls表示为分类损失。

4.根据权利要求1所述的基于随机帧补帧和注意力的视频交互动作检测方法,其特征是:金字塔特征采用6层Transformer层获得,每一层由LSTM、局部多头自注意力和MLP块交替层组成,在每个MSA或MLP之前应用LayerNorm,在每个块之后添加残差连接,通道 MLP,它有两个线性层,中间使用GELU激活,使用一个单步深度可分离1D卷积去实现下采样操作,模型为2倍下采样比率,具体公式如下：

$$\tilde{Z}^l = LSTM(Z^{l-1}) + Z^{l-1}, \quad l = 1...L$$
$$\bar{Z}^l = \alpha^l MSA(LN(\tilde{Z}^l)) + \tilde{Z}^l, \quad l = 1...L$$
$$\hat{Z}^l = \bar{\alpha}^l MLP(LN(\bar{Z}^l)) + \bar{Z}^l, \quad l = 1...L$$
$$Z^l = \downarrow(\hat{Z}^l), \quad l = 1...L$$

；

Z^{l-1}，$\tilde{Z}^l$，$\bar{Z}^l$，$\hat{Z}^l \in R^{T^{l-1} \times D}$，$Z^l \in R^{T^l \times D}$，$\alpha^l$和$\bar{\alpha}^l$是初始化为0的可学习的每通道缩放因子，$T^{l-1}$表示$l-1$层的时间序列长度；$T^l$表示$l$层的时间序列长度。

5.一种基于随机帧补帧和注意力的视频交互动作检测系统，其特征是：

包括特征提取模块，用于提取全局的时序信息；

时序自注意力模块，用于对全局的时序信息进行建模获得了包含多尺度局部信息的特征；

随机帧补帧数据增强模块，用于使原视频动作和边界清晰；

金字塔特征生成模块，用于将多尺度局部信息的特征通过多尺度的Transformer编码成6层的特征金字塔，并且将LSTM与Transformer进行结合；

分类模块，对每一个尺度的金字塔特征，分别输入到不同的1D卷积中来获得定位和分类的特征；

随机帧补帧数据增强模块中用到的公式如下：

原视频特征向量：$X = [x_1, x_2 ... x_T] \in R^{T*D}$，

T表示视频特征序列长度，D 表示特征维度；

把X分成t/k段：$X_{seg} = \left[\bar{x}_1, \bar{x}_2 ... \bar{x}_{\frac{T}{k}}\right]$，

每个$\bar{x}_i$包含k帧，i 表示第i个视频特征；从每个片段中随机取一帧，并复制k次，

$$X_{out} = \left[\varphi(\sigma(\bar{x}_1)), \varphi(\sigma(\bar{x}_2)) ... \varphi(\sigma(\bar{x}_{\frac{T}{k}}))\right] \in R^{T*D},$$

σ代表随机取帧，φ代表复制k次操作；

$$Loss = MSE(A_1, A_2)\theta,$$

A_1, A_2代表向量X和X_{out}经过backbone网络之后的新的特征向量，MSE均方损失函数，θ 表示调节系数，通常为1。

6.一种存储有计算机程序的计算机可读存储介质，其中，当所述计算机程序被处理器执行时，实现权利要求1至4中任一项所述的视频交互动作检测方法。

7.一种计算装置，包括：至少一个处理器；至少一个存储器，存储有计算机程序，当所述

计算机程序被所述至少一个处理器执行时,实现权利要求1至4中任一项所述的视频交互动作检测方法。

基于随机帧补帧和注意力的视频交互动作检测方法及系统

技术领域

[0001] 本发明属于计算机视觉和模式识别技术领域,尤其涉及一种基于随机帧补帧和注意力的视频交互动作检测方法及系统。

背景技术

[0002] 近几年中,随着深度学习技术的飞速发展,许多学者提出了许多基于深度学习技术的时序动作定位方法。及时识别动作实例并识别其类别,即时序动作定位,仍然是视频理解中的一个具有挑战性的问题。在TAL的深度模型开发方面取得了重大进展。以前的大多数工作都考虑使用动作Proposals[BMN]或Anchor窗口[GTAN],并为TAL开发了卷积神经网络[CDC,SSN]、循环神经网络[SS-TAD]和图神经网络[BC-GNN,G-TAD]。尽管在主要基准上取得了稳定的进展,但现有方法的准确性通常是以建模复杂性为代价的,包括越来越复杂的Proposal生成、Anchor设计和损失函数,网络结构和输出解码过程。同时,由于视频中动作边界不明确,现有的方法往往存在边界预测不准确的问题。

[0003] 如何解决时序动作定位的问题,在之前已经提出的方法中已经给出了一些解决方法,但是这些方法仍然存在着一些问题。基于Anchor的方法需要很强的先验知识,对每个数据集定义的anchor的数量也不一样,这些问题会影响最终的结果。虽然Actionness-Guided的方法能取得不错的效果,但是Actionness-Guided方法的计算量太大。因此Anchor-free的方法可能是一种很好的解决方案。

发明内容

[0004] 本发明的目的是解决时序动作定位问题,之前的时序动作定位方法要么需要对数据集很强的先验知识,要么计算量很大。本发明提出基于随机帧补帧和注意力的视频交互动作检测方法及系统,用于解决时序动作定位方法需要很强先验知识或者计算量很大的问题,通过全局与多尺度信息的聚合,时序位置关系的建模实现了对动作的精确定位,本发明方法识别精度高,从而为Anchor-free的时序动作定位问题提供了帮助。

[0005] 本发明解决技术问题的技术方案为:

[0006] 一种基于随机帧补帧和注意力的视频交互动作检测方法,包括以下步骤:

[0007] 步骤10.特征提取网络的选择

[0008] 选择基于Kinetics数据集预训练的I3D网络来进行特征的提取,将16个连续帧作为I3D的输入,使用步长为4的滑动窗口,在最后一个全连接层之前提取1024-D的特征,双流特征被进一步连接(2048-D)作为模型的输入;

[0009] 步骤20.自注意力全局信息建模

[0010] 在步骤10基础网络选择的基础上,对全局的时序信息进行建模,对于I3D网络的输出;用Polarized Self-Attention极化注意力来寻找帧与帧之间的关系并进行加权,通过这种基于自注意力的加权策略能够寻找到更重要的帧并赋予更高的权重;

[0011] 在Transformer网络之前加入1D卷,可以更好的合并局部上下文信息和稳定视觉

Transformer的训练，以此实现了全局信息的建模；

[0012] 步骤30. 随机帧补帧数据增强

[0013] 在第1步特征网络的输出上，通过将一个视频分为T/k个片段，从每个片段中随机取一帧，其余k-1帧与所取帧相同，来形成一个变化较大的新特征向量，相当于把视频给加速了，但是动作实际位置不变；

[0014] 把经过backbone的新特征向量与原视频特征向量计算一个mse损失，对它们进行约束，让它们拉近，互相学习一些信息，以此达到数据增强的目的；

[0015] 步骤40.金字塔特征的生成

[0016] 在步骤20步网络的基础上，将通过多尺度信息聚合模块之后的特征通过多尺度的Transformer编码成6层的特征金字塔，并且将LSTM与Transformer进行结合，将它们融合可以提供LSTM和Transformer模块所提供的补充历史信息和基于注意力的信息表示，提高了模型能力，还有就是能够解决单个模型在不同大小数据集上性能表现不同的问题，一般LSTM在小数据集上比Transformer表现更好，但Transformer在预训练后表现很突出；

[0017] 步骤50.边界定位与分类

[0018] 在得到6个尺度的金字塔特征之后；对每一个尺度的金字塔特征，分别输入到不同的1D卷积中来获得定位和分类的特征，之后采用分类特征来进行分类，采用定位特征进行边界的回归，在训练分类的过程中采用focal loss进行约束，在训练回归的过程中采用GIoU loss进行约束。

[0019] 上述的基于随机帧补帧和注意力的视频交互动作检测方法基础上，步骤30中公式如下：

[0020] 原视频特征向量：$X = [x_1, x_2 \dots x_T] \in R^{T*D}$；

[0021] 把X分成t/k段：$X_{seg} = \left[\overline{x}_1, \overline{x}_2 \dots \overline{x}_{\frac{T}{k}}\right]$，每个 $\overline{x}_i$ 包含k帧；

[0022] 从每个片段中随机取一帧，并复制k次，

[0023] $X_{out} = \left[\varphi(\sigma(\overline{x}_1)), \varphi(\sigma(\overline{x}_2)) \dots \varphi(\sigma(\overline{x}_{\frac{T}{k}}))\right] \in R^{T*D}$，$\sigma$ 代表随机取帧，φ代表复制k次操作；

[0024] $Loss = MSE(A_1, A_2)\theta$，$A_1, A_2$代表向量X和$X_{out}$经过backbone网络之后的新的特征向量，$MSE$均方损失函数。

[0025] 上述的基于随机帧补帧和注意力的视频交互动作检测方法基础上，在对于提取出来的特征通过Polarized Self-Attention中的Channel-only branch和Spatial-only branch进行操作，Channel-only branch定义如下：

[0026] $A^{ch}(X) = F_{SG}\left[W_{z|\theta_1}\left(\sigma_1(W_v(X)) \times F_{SM}(\sigma_2(W_q(X)))\right)\right], A^{ch}(X) \in R^{C\times 1\times 1}$

，其中 W_q、W_k、W_v是1×1卷积层，σ_1，σ_2是 reshape operator即把特征维度由

C/2×H×W改为C/2×HW，F_{SM}是softmax算子，X是矩阵点积运算$F_{SM}(X) = \sum_{j=1}^{N_p} \frac{e^{x_j}}{\sum_{m=1}^{N_p} e^{x_m}} x_j$，$W_v$、$W_q$和$W_z$之间的内部通道数是C/2，通道分支的输出是$Z^{ch} = A^{ch}(X) \odot^{ch} X \in R^{C\times H\times W}$，其中$\odot^{ch}$是通道乘法运算操作符；

[0027] Spatial-only branch定义如下：

$A^{sp}(X) = F_{SG}\left[\sigma_3\left(F_{SM}(\sigma_1(F_{GP}(W_q(X)))) \times \sigma_2(W_v(X))\right)\right]$，$A^{sp}(X) \in R^{1\times H\times W}$，其中$W_q$、$W_v$是标准的1×1卷积，$\sigma_1\sigma_2\sigma_3$是三个reshape operator，$F_{SM}$是softmax算子，$F_{GP}$是全局池化操作符，$F_{GP}(X) = \frac{1}{H\times W}\sum_{H}^{i=1}\sum_{w}^{j=1} X(:,i,j)$，空间分支的输出是$Z^{sp} = A^{sp}(X) \odot^{sp} X \in R^{C\times H\times W}$，其中$\odot^{sp}$是空间乘法运算操作符；

[0028] 通道分支和空间分支的输出在并行布局下组成：

[0029] $PSA_p(X) = Z^{ch} + Z^{sp} = A^{ch}(X) \odot^{ch} X + A^{sp}(X) \odot^{sp} X$。

[0030] 上述的基于随机帧补帧和注意力的视频交互动作检测方法基础上，每一个视频损失定义如下：$L = \sum_t \left(\frac{1}{T} L_{cls} + \frac{\lambda_{reg}}{T_+} 1_{ct} L_{reg}\right)$；

[0031] 其中T是输入序列的长度。1_{ct}是一个指示函数，表示时间步长t是否在动作范围内，即正样本，T_+是阳性样本总数，L应用于输出金字塔上的所有级别，并在训练期间对所有视频样本进行平均，λ_{reg}是一个平衡分类损失和回归损失的系数，L_{reg}用于距离回归的一个GIoU loss。

[0032] 上述的基于随机帧补帧和注意力的视频交互动作检测方法基础上，金字塔特征采用6层Transformer层获得，每一层由LSTM、局部多头自注意力和MLP块交替层组成，在每个MSA或MLP之前应用LayerNorm，在每个块之后添加残差连接，通道 MLP，它有两个线性层，中间使用GELU激活，使用一个单步深度可分离1D卷积去实现下采样操作，模型为2倍下采样比率，具体公式如下：

[0033]
$$\begin{aligned}
\tilde{Z}^l &= LSTM(Z^{l-1}) + Z^{l-1}, && l = 1\ldots L\\
\bar{Z}^l &= \alpha^l MSA(LN(\tilde{Z}^l)) + \tilde{Z}^l, && l = 1\ldots L\\
\hat{Z}^l &= \bar{\alpha}^l MLP(LN(\bar{Z}^l)) + \bar{Z}^l, && l = 1\ldots L\\
Z^l &= \downarrow(\hat{Z}^l), && l = 1\ldots L
\end{aligned},$$

Z^{l-1}，$\tilde{Z}^{l}$，$\bar{Z}^{l}$，$\hat{Z}^{l} \in R^{T^{l-1} \times D}$，$Z^{l} \in R^{T^{l} \times D}$，$\alpha^{l}$和$\bar{\alpha}^{l}$是初始化为0的可学习的每通道缩放因子，$T^{l-1}/T^{l}$是下采样比例。

[0034] 本发明实施例中，还提供了一种基于随机帧补帧和注意力的视频交互动作检测系统，包括特征提取模块，用于提取全局的时序信息；时序自注意力模块，用于对全局的时序信息进行建模获得了包含多尺度局部信息的特征；随机帧补帧数据增强模块，用于使原视频动作和边界清晰；金字塔特征生成模块，用于将多尺度局部信息的特征通过多尺度的Transformer编码成6层的特征金字塔，并且将LSTM与Transformer进行结合；分类模块，对每一个尺度的金字塔特征，分别输入到不同的1D卷积中来获得定位和分类的特征。

[0035] 本发明实施例中，还提供了一种存储有计算机程序的计算机可读存储介质，其中，当所述计算机程序被处理器执行时，实现所述的视频交互动作检测方法。

[0036] 本发明实施例中，还提供了一种计算装置，包括：至少一个处理器；至少一个存储器，存储有计算机程序，当所述计算机程序被所述至少一个处理器执行时，实现所述的视频交互动作检测方法。

[0037] 发明内容中提供的效果仅仅是实施例的效果，而不是发明所有的全部效果，上述技术方案具有如下优点或有益效果：

[0038] 1)通过自注意力机制能够寻找到更重要的帧并赋予更高的权重来实现全局信息的建模。

[0039] 2)通过对原视频特征进行随机帧补帧，使原视频变化更大，以此来达到数据增强。

[0040] 3)通过将LSTM和Transformer结合，提高了模型能力，解决单个模型在不同大小数据集上性能表现不同的问题。

附图说明

[0041] 附图用来提供对本发明的进一步理解，并且构成说明书的一部分，与本发明的实施例一起用于解释本发明，并不构成对本发明的限制。

[0042] 图1为本发明的结构图。

具体实施方式

[0043] 为了能清楚说明本方案的技术特点，下面通过具体实施方式，并结合其附图，对本发明进行详细阐述。

[0044] 实施例1 如图1所示，为本发明的一种基于随机帧补帧和注意力的视频交互动作检测方法的操作流程图，该方法包括以下步骤：

[0045] 步骤10.特征提取网络的选择

[0046] 在时序动作定位任务中，需要首先选取优秀的特征提取器来获得鲁棒的特征，由于时序动作定位任务的特性，必须要选取能够提取时序信息的特征提取器。因此本文采用了双流的I3D网络来进行特征的提取。RGB流的输入为连续的视频帧，能够同时提取到时间和空间特征，对于Flow流，输入为连续的光流帧，能够进一步对时序信息进行提取和建模；选择基于Kinetics数据集预训练的I3D网络来进行特征的提取，将16个连续帧作为I3D的输入，使用步长为4的滑动窗口，在最后一个全连接层之前提取1024-D的特征，双流特征被进

一步连接(2048-D)作为模型的输入；

[0047] 步骤20.自注意力全局信息建模

[0048] 在步骤10基础网络选择的基础上，对全局的时序信息进行建模，对于I3D网络的输出；用Polarized Self-Attention极化注意力来寻找帧与帧之间的关系并进行加权，通过这种基于自注意力的加权策略能够寻找到更重要的帧并赋予更高的权重；

[0049] 在Transformer网络之前加入1D卷，可以更好的合并局部上下文信息和稳定视觉Transformer的训练，以此实现了全局信息的建模；

[0050] 步骤30. 随机帧补帧数据增强

[0051] 为在未剪辑的视频当中通常包含不相干活动的背景，导致动作边界是不清楚的；为了扩大视频的变化，使得边界更加的明显，提出了随机帧补帧用来数据增强；

[0052] 在第1步特征网络的输出上，通过将一个视频分为T/k个片段，从每个片段中随机取一帧，其余k-1帧与所取帧相同，来形成一个变化较大的新特征向量，相当于把视频给加速了，但是动作实际位置不变；

[0053] 把经过backbone的新特征向量与原视频特征向量计算一个mse损失，对它们进行约束，让它们拉近，互相学习一些信息，以此达到数据增强的目的；

[0054] 步骤40.金字塔特征的生成

[0055] 在步骤20步网络的基础上，将通过多尺度信息聚合模块之后的特征通过多尺度的Transformer编码成6层的特征金字塔，并且将LSTM与Transformer进行结合，将它们融合可以提供LSTM和Transformer模块所提供的补充历史信息和基于注意力的信息表示，提高了模型能力，还有就是能够解决单个模型在不同大小数据集上性能表现不同的问题，一般LSTM在小数据集上比Transformer表现更好，但Transformer在预训练后表现很突出；

[0056] 步骤50.边界定位与分类

[0057] 通过步骤40得到金字塔特征，在得到金字塔特征之后，分类头检查金字塔所有 L 层的每个时刻t，并预测每个时刻t的动作 $p(a_t)$ 的概率，这个头是使用连接到每个金字塔层的轻量级1D卷积网络来实现的，参数在所有级别都是共享的；分类网络使用3层核大小为3的1D卷积、层归一化(前2层)和ReLU激活来实现；在每个输出维度上附加一个sigmoid函数来预测C个动作类别的概率；回归头类似于分类头，回归头检查金字塔上所有L层的每一刻t；

[0058] 不同之处在于回归头预测到动作开始和偏移的距离 (d_t^s, d_t^e) ，仅当当前时间步长t位于动作中，每个金字塔级别都预先指定了输出回归范围，回归头同样是使用一维卷积网络采用与分类网络相同的设计，只是在末端附加了一个ReLU用于距离估计；模型对于每个时候t输出的$(p(a_t), d_t^s, d_t^e)$，包括动作类别$p(a_t)$的概率和到动作边界的距离(d_t^s, d_t^e)；损失函数同样遵循极简设计，只有两项(1)L_{cls}一个focal loss对于C类二分类；(2)L_{reg}用于距离回归的一个GIoU loss；

[0059] 步骤60时序动作定位效果

[0060] 在 THUMOS14数据集上，使用在Kinetics上预训练的双流I3D在THUMOS14上提取视频特征；将16个连续帧作为I3D的输入，使用步幅为4的滑动窗口，在最后一个全连接层之前

提取1024-D的特征;两个流特征被进一步连接(2048-D)作为模型的输入;mAP@[0.3:0.1:0.7]被用来评估本发明的模型。本发明训练了50个epoch,其中线性热身为5个epoch;初始学习率为1e-4,使用余弦学习率衰减;小批大小为2,权重衰退为1e-4;在消融的基础上,局部自我注意的窗口大小为19;还结合了来自UntrimmedNet的外部分类评分;对于ActivityNet1.3 数据集,使用双流I3D进行特征提取,但将滑动窗口的步长增加到16;将提取的特征通过线性插值下采样到固定长度128;为了进行评估,使用mAP@[0.5:0.05:0.95],并报告了平均mAP;模型训练了15个周期,其中线性热身为5个周期;学习率为1e-3,小批次大小为16,权重衰退为1e-4。窗口大小为25用于局部自我关注;此外,结合了外部分类结果,类似地,本发明考虑来自TSP的预训练方法,并将模型与同一组baseline进行比较,包括最接近的竞争对手单阶段模型。

[0061] 在测试过程中,在推理时,将完整的序列输入到模型中,因为模型中没有使用位置嵌入。我们的模型取输入视频X,并输出$\{(p(a_t), d_t^s, d_t^e)\}$在所有金字塔层的每个时间步长$t$。每个时间步$t$进一步解码一个动作实例$(e_t = t - d_t^s, s_t = t + d_t^e, p(a_t))$。$e_t$ 和 s_t 是动作的开始和偏移,$p(a_t)$是动作置信度评分。使用Soft-NMS进一步处理结果动作候选,以删除高度重叠的实例,从而得到动作的最终输出。

[0062] 对于本发明的实验效果与其他方法在THUMOS14数据集和ActivityNet1.3数据集的比较在下表中:

[0063]

方法	THUMOS14						ActivityNet1.3			
	0.3	0.4	0.5	0.6	0.7	Avg.	0.5	0.75	0.95	Avg.
BMN	56.0	47.4	38.8	29.7	20.5	38.5	50.1	34.8	8.3	33.9
G-TAD	54.5	47.6	40.3	30.8	23.4	39.3	50.4	34.6	9.0	34.1
A2Net	58.6	54.1	45.5	32.5	17.2	41.6	43.6	28.7	3.7	27.8
TCANet	60.6	53.2	44.6	36.8	26.7	44.3	52.3	36.7	6.9	35.5
RTD-Net	68.3	62.3	51.9	38.8	23.7	49.0	47.2	30.7	8.6	30.8
VSGN	66.7	60.4	52.4	41.0	30.4	50.2	52.4	36.0	8.4	35.1
ContextLoc	68.3	63.8	54.3	41.8	26.2	50.9	56.0	35.2	3.6	34.2
AFSD	67.3	67.3	62.4	55.5	43.7	31.1	52.4	35.3	6.5	34.4
TadTR	74.8	69.1	60.1	46.6	32.8	56.7	49.1	32.6	8.5	32.3
ActionFormer	82.1	77.8	71.0	59.4	43.9	66.8	54.7	37.8	8.4	36.6
本发明方法	83.1	79.2	72.7	60.9	45.6	68.3	54.4	37.3	8.3	36.2

[0064] 在THUMOS14数据集上,本发明取得了最好的效果,当计算tIoU从0.3～0.7的平均mAP时,取得了68.3的效果,在ActivityNet1.3数据集上,虽然本发明没有取得最好的效果,但是取得的效果超过了绝大多数方法,当计算tIoU从0.5～0.95的平均mAP时,取得的36.18的效果仍然是一个很好的效果。

[0065] 本实施例中,在对于提取出来的特征通过Polarized Self-Attention中的Channel-only branch和Spatial-only branch进行操作。Channel-only branch定义如下:

[0066] $A^{ch}(X) = F_{SG}\left[W_{z|\theta_1}\left(\sigma_1(W_v(X)) \times F_{SM}(\sigma_2(W_q(X)))\right)\right], A^{ch}(X) \in R^{C\times 1\times 1}$,其中 W_q、W_k、W_v是1×1卷积层，σ_1， σ_2是 reshape operator即把特征维度由C/2×H×W改为C/2×HW，F_{SM} 是 softmax 算子，X是矩阵点积运算 $F_{SM}(X) = \sum_{j=1}^{N_p} \frac{e^{x_j}}{\sum_{m=1}^{N_p} e^{x_m}} x_j$ ，W_v、W_q和W_z之间的内部通道数是C/2，通道分支的输出是$Z^{ch} = A^{ch}(X) \odot^{ch} X \in R^{C\times H\times W}$，其中$\odot^{ch}$是通道乘法运算操作符；其原理为：先用卷积核为1的一维卷积将输入的特征X转换成了Q和V，其中Q的通道被完全压缩，而V的通道维度依旧保持在一个比较高的水平(也就是C/2)，因为Q的通道维度被压缩，如上面所说的那样，就需要通过HDR进行信息的增强，因此用Softmax对Q的信息进行了增强，然后将Q和K进行矩阵乘法，并在后面接上卷积核为1的一维卷积、LN将通道上C/2的维度升为C，最后用Sigmoid函数使得所有的参数都保持在0-1之间。

[0067] Spatial-only branch定义如下：

[0068] $A^{sp}(X) = F_{SG}\left[\sigma_3\left(F_{SM}(\sigma_1(F_{GP}(W_q(X)))) \times \sigma_2(W_v(X))\right)\right],\ A^{sp}(X) \in R^{1\times H\times W}$ ，其中 W_q、W_v 是标准的1 × 1卷积，$\sigma_1\sigma_2\sigma_3$是三个 reshape operator，F_{SM} 是 softmax算子，F_{GP}是全局池化操作符，$F_{GP}(X) = \frac{1}{H\times W}\sum_H^{i=1}\sum_w^{j=1} X(:,i,j)$，空间分支的输出是 $Z^{sp} = A^{sp}(X) \odot^{sp} X \in R^{C\times H\times W}$ ，其中$\odot^{sp}$是空间乘法运算操作符；可以看出，与Channel-only branch相似，先用了卷积核为1的一维卷积将输入的特征转换为了Q和V，其中，对于Q特征，还用了GlobalPooling对时间维度压缩，转换成了1的大小；而V特征的时间维度则保持在一个较大的水平；由于Q的时间维度被压缩了，所以就用了Softmax对Q的信息进行增强；然后将Q和K进行矩阵乘法，然后接上reshape和sigmoid使得所有的参数都保持在0-1之间。

[0069] 通道分支和空间分支的输出在并行布局下组成：

[0070] $PSA_p(X) = Z^{ch} + Z^{sp} = A^{ch}(X) \odot^{ch} X + A^{sp}(X) \odot^{sp} X$ ，把增强后含有全局信息的特征经过一个浅层的卷积神经网络对于时间序列数据更好的合并局部上下文信息和稳定视觉Transformer的训练是有帮助的。

[0071] 本实施例中，步骤30中公式如下：

[0072] 原视频特征向量：$X = [x_1, x_2 \ldots x_T] \in R^{T*D}$；

[0073] 把X分成t/k段：$X_{seg} = \left[\overline{x}_1, \overline{x}_2 \ldots \overline{x}_{\frac{T}{k}}\right]$，每个 $\overline{x}_i$ 包含k帧；

[0074] 从每个片段中随机取一帧，并复制k次，

[0075] $X_{out}=\left[\varphi(\sigma(\bar{x}_1)),\varphi(\sigma(\bar{x}_2))\ldots\varphi(\sigma(\bar{x}_{\frac{T}{k}}))\right]\in R^{T*D}$，$\sigma$ 代表随机取帧，φ 代表复制k次操作；

[0076] $Loss=MSE(A_1,A_2)\theta$，A_1,A_2代表向量X和X_{out}经过backbone网络之后的新的特征向量，MSE均方损失函数。

[0077] 本实施例中，每一个视频损失定义如下：$L=\sum_t\left(\frac{1}{T}L_{cls}+\frac{\lambda_{reg}}{T_+}1_{ct}L_{reg}\right)$；其中 T 是输入序列的长度。1_{ct}是一个指示函数，表示时间步长t是否在动作范围内，即正样本，T_+ 是阳性样本总数，L 应用于输出金字塔上的所有级别，并在训练期间对所有视频样本进行平均，λ_{reg} 是一个平衡分类损失和回归损失的系数，λ_{reg} 用于距离回归的一个GIoU loss。

[0078] 本实施例中，金字塔特征采用6层Transformer层获得，每一层由LSTM、局部多头自注意力和MLP块交替层组成，在每个MSA或MLP之前应用LayerNorm，在每个块之后添加残差连接，通道 MLP，它有两个线性层，中间使用GELU激活，使用一个单步深度可分离1D卷积去实现下采样操作，模型为2倍下采样比率，具体公式如下：

[0079]
$$\begin{aligned}
\tilde{Z}^l &= LSTM(Z^{l-1})+Z^{l-1}, && l=1\ldots L\\
\bar{Z}^l &= \alpha^l MSA(LN(\tilde{Z}^l))+\tilde{Z}^l, && l=1\ldots L\\
\hat{Z}^l &= \bar{\alpha}^l MLP(LN(\bar{Z}^l))+\bar{Z}^l, && l=1\ldots L\\
Z^l &= \downarrow(\hat{Z}^l), && l=1\ldots L
\end{aligned}$$

Z^{l-1}，$\tilde{Z}^l$，$\bar{Z}^l$，$\hat{Z}^l\in R^{T^{l-1}\times D}$，$Z^l\in R^{T^l\times D}$，$\alpha^l$和$\bar{\alpha}^l$是初始化为0的可学习的每通道缩放因子，$T^{l-1}/T^l$是下采样比例。

[0080] 实施例2 本发明实施例中，还提供了一种基于随机帧补帧和注意力的视频交互动作检测系统，包括特征提取模块，用于提取全局的时序信息；时序自注意力模块，用于对全局的时序信息进行建模获得了包含多尺度局部信息的特征；随机帧补帧数据增强模块，用于使原视频动作和边界清晰；金字塔特征生成模块，用于将多尺度局部信息的特征通过多尺度的Transformer编码成6层的特征金字塔，并且将LSTM与Transformer进行结合；分类模块，对每一个尺度的金字塔特征，分别输入到不同的1D卷积中来获得定位和分类的特征。

[0081] 实施例3 本发明实施例中，还提供了一种存储有计算机程序的计算机可读存储介质，其中，当所述计算机程序被处理器执行时，实现所述的视频交互动作检测方法。

[0082] 实施例4 本发明实施例中，还提供了一种计算装置，包括：至少一个处理器；至少

一个存储器，存储有计算机程序，当所述计算机程序被所述至少一个处理器执行时，实现所述的视频交互动作检测方法。

[0083] 上述虽然结合附图对发明的具体实施方式进行了描述，但并非对本发明保护范围的限制，在本发明的技术方案的基础上，本领域技术人员不需要付出创造性劳动即可做出的各种修改或变形仍在本发明的保护范围以内。

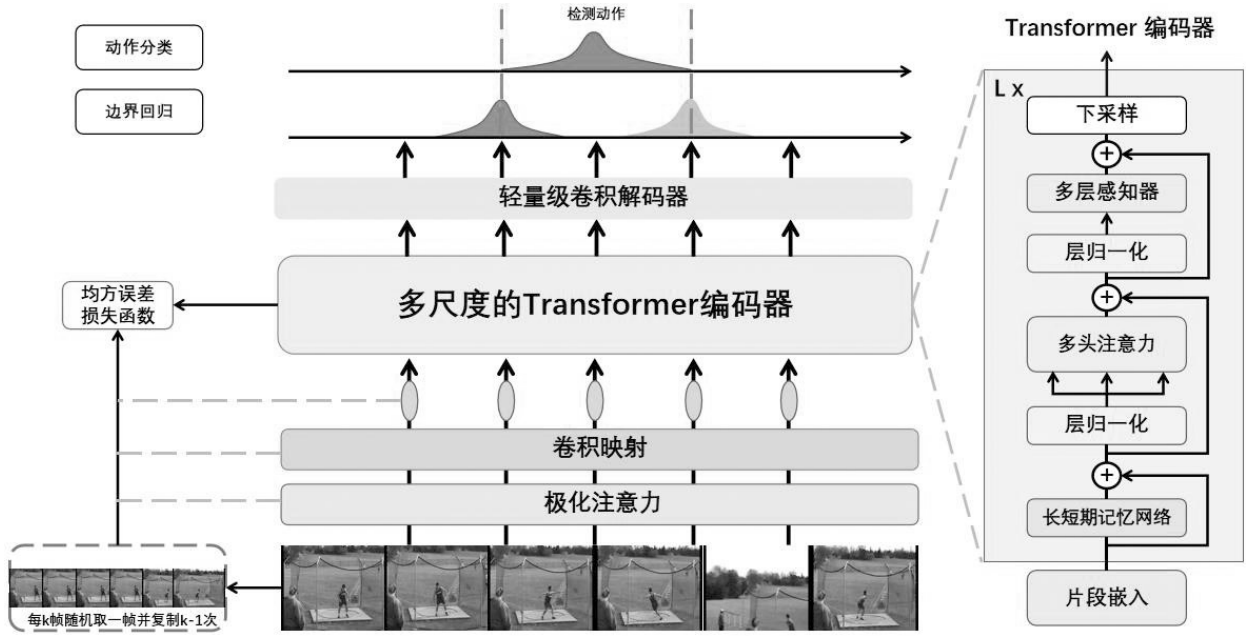

检测动作
动作分类
边界回归
轻量级卷积解码器
多尺度的Transformer编码器
均方误差
损失函数
卷积映射
极化注意力
每k帧随机取一帧并复制k-1次
Transformer 编码器
L x
下采样
多层感知器
层归一化
多头注意力
层归一化
长短期记忆网络
片段嵌入

图 1