计算机科学
COMPUTER SCIENCE

基于改进双流视觉Transformer的行为识别模型

雷永升, 丁锰, 沈尧, 李居昊, 赵东越, 陈福仕

引用本文

雷永升, 丁锰, 沈尧, 李居昊, 赵东越, 陈福仕. 基于改进双流视觉Transformer的行为识别模型[J]. 计算机科学, 2024, 51(7): 229-235.

LEI Yongsheng, DING Meng, SHEN Yao, LI Juhao, ZHAO Dongyue, CHEN Fushi. Action Recognition Model Based on Improved Two Stream Vision Transformer [J]. Computer Science, 2024, 51(7): 229-235.

相似文章推荐（请使用火狐或 IE 浏览器查看文章）

Similar articles recommended (Please use Firefox or IE to view the article)

多粒度空间注意力与空间先验监督的DETR

DETR with Multi-granularity Spatial Attention and Spatial Prior Supervision

计算机科学, 2024, 51(6): 239-246. https://doi.org/10.11896/jsjkx.230300218

基于3D骨架相似性的自适应移位图卷积神经网络人体行为识别算法

Human Action Recognition Algorithm Based on Adaptive Shifted Graph Convolutional Neural Network with 3D Skeleton Similarity

计算机科学, 2024, 51(4): 236-242. https://doi.org/10.11896/jsjkx.221200120

基于Depth-wise卷积和视觉Transformer的图像分类模型

Novel Image Classification Model Based on Depth-wise Convolution Neural Network andVisual Transformer

计算机科学, 2024, 51(2): 196-204. https://doi.org/10.11896/jsjkx.221100234

Transformer在计算机视觉场景下的研究综述

Review of Transformer in Computer Vision

计算机科学, 2023, 50(12): 130-147. https://doi.org/10.11896/jsjkx.221100076

多流融合的轻量级图卷积行为识别算法

Lightweight Graph Convolution Action Recognition Algorithm Based on Multi-streamFusion

计算机科学, 2023, 50(11A): 220800147-6. https://doi.org/10.11896/jsjkx.220800147

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230500054

基于改进双流视觉 Transformer 的行为识别模型

雷永升[1]　丁　锰[1,2]　沈　尧[1]　李居昊[1]　赵东越[1]　陈福仕[1]
1 中国人民公安大学侦查学院　北京 100038
2 中国人民公安大学公共安全行为科学实验室　北京 100038
(834624067@qq.com)

摘　要　针对现有行为识别方法中抗背景干扰能力差和准确率低等问题,提出了一种改进的双流视觉 Transformer 行为识别模型。该模型采用分段采样的方法来增加模型对长时序列数据的处理能力;在网络头部嵌入无参数的注意力模块,在降低动作背景干扰的同时,增强了模型的特征表示能力;在网络尾部嵌入时间注意力模块,通过融合时域高语义信息来充分提取时序特征。文中提出了一种新的联合损失函数,旨在增大类间差异并减少类内差异;采用决策融合层以充分利用光流与 RGB 流特征。针对上述改进模型,在基准数据集 UCF101 和 HMDB51 上进行消融及对比实验,消融实验结果验证了所提方法的有效性,对比实验结果表明,所提方法相比时间分段网络在两个数据集上的准确率分别提高了 3.48%和 7.76%,优于目前的主流算法,具有较好的识别效果。

关键词:行为识别;视觉 Transformer;SimAM 无参注意力;时间注意力;联合损失
中图分类号　TP391.7

Action Recognition Model Based on Improved Two Stream Vision Transformer

LEI Yongsheng[1],DING Meng[1,2],SHEN Yao[1],LI Juhao[1],ZHAO Dongyue[1] and CHEN Fushi[1]
1 Department of Criminal Investigation,People's Public Security University of China,Beijing 100038,China
2 Public Security Behavioral Science Lab,People's Public Security University of China,Beijing 100038,China

Abstract　To address the issues of poor resistance to background interference and low accuracy in existing action recognition methods,an improved dual stream visual Transformer action recognition model is proposed. The model adopts a segmented sampling method to increase its processing ability for long-term sequence data; embedding a parameter free attention module in the network header enhances the model's feature representation ability while reducing action background interference;embedding a temporal attention module at the tail of the network to fully extract temporal features by integrating high semantic information in the time domain. A new joint loss function is proposed in the paper,aiming to increase inter class differences and reduce intra class differences. Adopting a decision fusion layer to fully utilize the features of optical flow and RGB flow. In response to the above improved model,comparative and ablation experiments are conducted on the benchmark datasets UCF101 and HMDB51. The ablation experiment results verify the effectiveness of the proposed method. The comparison results show that the accuracy of the proposed method is 3.48% and 7.76% higher than that of the time segmented network on the two datasets,respectively,which is better than the current mainstream algorithms and has good recognition performance.

Keywords　Action recognition,Vision Transformer,SimAM parameter-free attention,Temporal attention,Joint loss

1　引言

近年来,行为识别技术的应用领域越来越广泛,如智能监控[1]、自动驾驶[2]、异常检测[3]、智慧教育[4]等。行为识别是深度学习领域的一个重要研究方向,其目的是通过计算机视觉技术和机器学习方法,从图像和视频数据中自动检测、识别和预测人类和其他实体的动作和行为。

深度学习技术,尤其是卷积神经网络(CNN)和视觉 Transformer(ViT),在行为识别任务中已经表现出卓越的性能。在将深度学习方法应用于行为识别的早期阶段,Karpathy 等[5]利用 CNN 直接从堆叠的视频帧中提取时空特征,并进行端到端的训练,但识别精度远不及改进密集轨迹算法(iDT)等传统方法。视频数据的本质是在图像数据的基础上增加时间维度,而若想进行正确的分类,还需要在提取空间特征

到稿日期:2023-05-09　返修日期:2023-10-09
基金项目:公安学一流学科培优行动及公共安全行为科学实验室建设项目(2023ZB02)
This work was supported by the First-class Discipline Training Program for Public Security Studies and Construction Project for Laboratory of Public Safety Behavior Science(2023ZB02).
通信作者:丁锰(dingmeng@ppsuc.edu.cn)

的基础上有效地提取时间特征。光流法[6]是利用图像序列中相邻帧像素之间的变化，从而计算出相邻帧物体运动信息，是能够有效提取时间特征的方法。Simonyan 等[7]基于光流信息设计了双流网络，将视频帧和堆叠的光流图输入两个 CNN，分别提取空间特征和时间特征，取得了与 iDT 相当的识别效果，从而验证了光流法对提取行为时序特征的有效性。时间分段网络(TSN)基于双流网络使用了稀疏采样策略，实现了长期特征捕获，能够识别更多复杂动作，降低了信息冗余，以较低代价实现了端到端的学习。长期循环神经网络(LRCN)基于双流网络添加了长短时记忆网络(LSTM)，以提高 CNN 的长时表征能力，但 LSTM 本身训练困难，且由于时序先后的严格迭代，从而导致其训练效率低下。Neimark 等[8]先利用 CNN 提取空间特征，再通过 Transformer 提取时间特征，由于 Transformer 中的自注意力机制使其能够直接获取全局信息，因此提高了模型长时表征能力，此外，由于训练是并行的，其效率更高。Arnab 等[9]采用了纯 ViT 的架构，在使用 ViT 提取空间特征后输入 Transformer 进行分类。Fan 等[10]关注到了行为细粒度，构建特征金字塔抽取多尺度行为特征并进行融合，虽然采用金字塔型多尺度处理表现良好，但下采样部分会导致时空信息流失。Yan 等[11]采用多尺度时间段对视频进行编码，再采用横向连接融合多尺度时间段特征，从而避免了部分时空信息的丢失。Bertasius 等[12]在 Transformer 的每层中都加入了时间注意力块，以实现对时序特征的提取。这些方法虽然在 Kinetics 数据集上有很好的表现，但模型都较深且较大，这不仅导致计算量大，而且在小数据集上很容易出现过拟合的情况。

目前，主流的行为识别方法已经逐渐转向纯 Transformer 的架构，但其中仍然存在许多问题，如需要大量的标注数据，难以在小数据集上进行训练；对背景干扰的问题关注不足；对于长时间序列和多模态数据的处理能力不足等。为了解决这些问题，本文提出了一种基于预训练 ViT 的时空双流网络模型。该模型以 ViT 为基础，通过迁移学习的方法针对行为识别任务进行了改进。迁移学习是一种通过已经训练好的模型来进行新任务训练的方法，这种方法可以减少过拟合现象以及模型对标注数据的依赖，从而提高模型的泛化能力。

本文还采用了其他的技术手段来优化模型的性能，包括嵌入无参型注意力机制[13](SimAM)来减少动作背景的干扰，采用分段采样的方法来处理长时序列，融合光流数据与 RGB 数据信息，将双流网络同视觉自注意力机制相结合，在降低计算复杂度的同时，提高了模型的准确率。

2 网络结构

本文模型主要由 4 个部分组成：基于 SimAM 的注意力机制模块、基于预训练 ViT 的时空特征提取模块、时间注意力模块、决策融合模块。此外，还对损失函数进行了改进。

如图 1 所示，首先，对视频进行前处理得到 n 段 RGB 帧与连续光流帧；其次，使用基于 SimAM 的注意力机制模块为 RGB 和光流图像赋予权重，并将它们输入基于预训练 ViT 的时空特征提取模块进行特征提取；然后，使用时间注意力模块提取时序信息并分段融合；最后，将两个模态的结果在决策融合层进行融合，从而得到最终的分类结果。

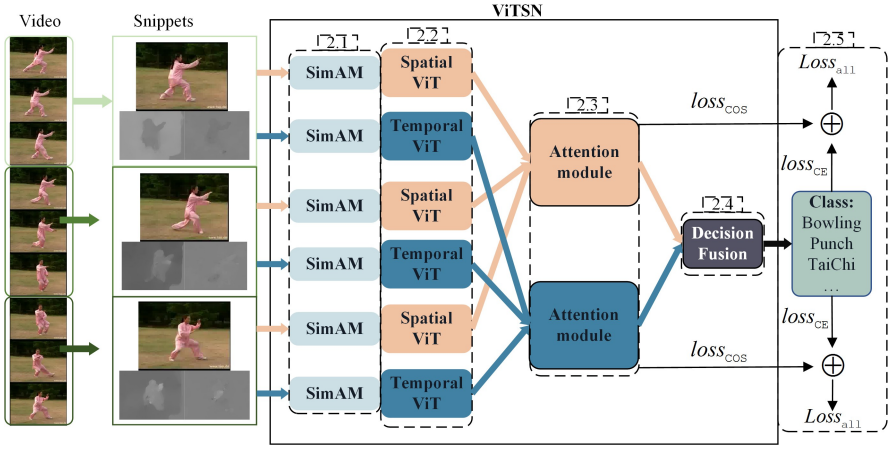

图 1 基于 ViT 的时空双流网络模型

Fig. 1 Spatiotemporal two stream network model based on ViT

2.1 SimAM 注意力机制

为了提高模型的抗干扰能力，本文提出的改进模型在骨干网络中引入了 3D 的无参型 SimAM 注意力机制。信息丰富的神经元通常会抑制周围神经元，即产生空域抑制，为了赋予具有空域抑制效应的神经元更高的权重，本文定义每个神经元的能量函数，测量目标神经元与其他神经元的线性可分性，从而找到重要的神经元，并根据重要性对其加权。能量函数定义如式(1)所示：

$$e_t^* = \frac{4(\hat{\sigma}^2+\lambda)}{(t-\hat{\mu})^2+2\hat{\sigma}^2+2\lambda} \tag{1}$$

其中，t 为目标神经元，e_t^* 表示目标神经元的能量，$\hat{\mu}$ 与 $\hat{\sigma}^2$ 分别为单个通道中除指定神经元外其他所有神经元的均值与方差，λ 为超参数。式(2)通过定义线性可分函数实现对模型各层神经元的评估，其中 $\boldsymbol{E}$ 为 e_t^* 组成的矩阵，sigmoid 为激活函数，目的是引入非线性。目标神经元能量越低，其与相邻神经元的区分度越高，则重要性也就越强，再根据重要性对目标神经元进行加权。

$$\tilde{\boldsymbol{X}} = \mathrm{sigmoid}\left(\frac{1}{\boldsymbol{E}}\right) \odot X \tag{2}$$

相比一维的挤压和激励网络(SENet)与二维的卷积注意力

机制(CBAM)，三维的 SimAM 注意力机制通过计算所有神经元的重要性，能够同时关注通道维度与空间维度。其能量函数的闭式解有效地评估了骨干网络在提取特征方面的重要性，在不增加额外的可学习参数的情况下，能够更高效更全面地评估特征的权重，从而减弱背景干扰，增强动作特征，提高模型的抗干扰能力。

2.2 基于 ViT 的特征提取

基于 ViT 的特征提取网络结构如图 2 所示。首先对图像切块并将所有块映射为一维向量，在加入类别 token 及位置编码后，输入 Transformer Encoder，该编码器共 12 层，每层由多头自注意力机制、层归一化、Dropout 和多层感知机组成。

多头自注意力机制的作用类似于 CNN 的多个卷积核，不同注意力学习关注不同的信息，帮助网络捕捉更丰富的特征；层归一化通过消除不同层之间的输入分布差异，来加速模型的收敛；Dropout 通过随机丢弃网络中部分神经元来防止过拟合，提高泛化能力。编码层同时利用了底层的高分辨率特征以及高层的高语义特征，充分捕捉图像块特征之间的空间关系，从而得到更具有表征性的图像特征向量。

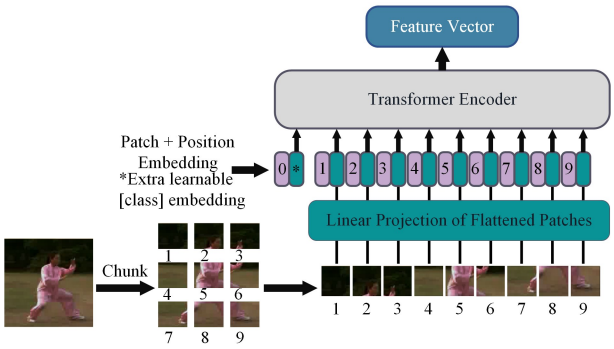

图 2 基于 ViT 的特征提取网络结构

Fig. 2 Feature extraction network structure based on ViT

本文中的双流特征提取网络以 ViT 网络模型为基础，但两个流的输入数据不同。空间网络的输入为 n 帧 RGB 图像，通过提取静态图像的特征并沿时间维度融合不同帧的特征，以对视频中的动作进行分类识别。时间网络则以 n 段连续的 10 帧光流图作为输入，光流图是相邻帧像素在时间域上的变化，其包括水平矢量和垂直矢量两个通道的特征，能够反映运动信息，有利于提取动作信息。

2.3 时间自注意力机制

本文使用了时间自注意力机制，这种机制可以让模型捕捉时间特征的变化，从而进一步优化时序特征的提取，具体实现方式为：在使用 ViT 充分提取空间特征后，对不同帧的相同空间编号块进行时间自注意力运算。如图 3 所示，对第 T 帧的蓝色块与第 $T-\delta$、$T+\delta$ 帧(δ 为分段间隔)的绿色块分别进行自注意力运算，是时间自注意力机制；而对第 T 帧中的所有红色块进行自注意力运算，则是空间自注意力机制。

在充分提取空间特征后，将形状为 $3\times196\times768$ 的输出矩阵进行变形，得到形状为 $196\times3\times768$ 的矩阵 $\boldsymbol{X}$。将矩阵 $\boldsymbol{X}$ 进行线性投影转化为 3 个不同的矩阵，即 $\boldsymbol{Q}$ 查询矩阵、$\boldsymbol{K}$ 键矩阵、$\boldsymbol{V}$ 值矩阵，$\boldsymbol{Q}$ 的维度与投影前的输入维度相同，自注意力机制的计算式如下：

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}\cdot\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_K}}\right)\cdot\boldsymbol{V} \tag{3}$$

其中，d_K 是 $\boldsymbol{K}$ 矩阵的行向量维度，作用是防止 Softmax 的输入过大。对于输入的矩阵 $\boldsymbol{X}$，整个时间自注意力的过程如下：

$$\boldsymbol{X}_1=Attention(LayerNorm(\boldsymbol{X})) \tag{4}$$

$$\boldsymbol{X}_2=\boldsymbol{X}_1+Dropout(\boldsymbol{X}_1) \tag{5}$$

$$\boldsymbol{X}_3=MLP(LayerNorm(\boldsymbol{X}_2)) \tag{6}$$

$$Output=\boldsymbol{X}_3+Dropout(\boldsymbol{X}_3) \tag{7}$$

其中，$LayerNorm()$为正则化层，旨在保证数据特征分布的稳定性；$MLP()$为多层感知机，通过对特征向量进行缩放来提取更抽象和有意义的特征表示；$Dropout()$为一种正则化技术，通过随机删除网络中的某些神经元来减轻模型的过拟合现象。

(a)frame $T-\delta$ (b)frame T (c)frame $T+\delta$

图 3 时间自注意力机制(电子版为彩图)

Fig. 3 Temporal self-attention mechanism

2.4 决策层融合

空间网络可以提取视频中的颜色和纹理信息，时间网络可以捕捉视频中的运动信息。然而，在实际应用中，视频数据往往包含噪声等干扰因素，这些因素可能会影响模型的性能和准确性。通过使用决策层融合，将不同来源的信息进行整合，综合利用它们的优点可以提高模型的鲁棒性和抗干扰能力，减少因无关信息影响而导致的错误。

通过空间网络和时间网络分别提取特征，得到两个长度为类别数的概率向量，根据权值将两个向量相加，如式(8)所示：

$$\boldsymbol{V}_F=\alpha\times\boldsymbol{V}_S+(1-\alpha)\times\boldsymbol{V}_T \tag{8}$$

其中，$\boldsymbol{V}_S$ 代表空间网络的概率向量，$\boldsymbol{V}_T$ 代表时间网络的概率向量，$\boldsymbol{V}_F$ 代表融合后的向量。在确定权重值时，按照 0.1 的步长在 0 到 1 之间对 α 进行调整，并根据准确率最大化来确定最优的权重值。

2.5 联合损失函数

交叉熵损失函数常用于分类问题，它刻画了两个概率分布之间的距离，通常用交叉熵来衡量模型输出的概率分布与真实标签的概率分布之间的差异。对于单个样本，假设 y 为真实分布，$\hat{y}$ 为模型输出分布，则两个分布之间差异的计算式为：

$$loss_{\mathrm{CE}}=-\sum_{i=1}^{n}y_i\ln\hat{y}_i \tag{9}$$

交叉熵损失函数使用常规的反向传播算法，将模型输出与真实标签之间的差值作为误差反向传递，从而计算出每个参数的梯度。

余弦相似度损失函数常用于度量样本在特征空间中的相似度。在实验中，由于输入为同视频的多段帧，因此可以使用最小化余弦相似度损失函数来训练模型，从而使得同类别的多段帧在特征空间中更加接近，不同类别的多段帧在特征

空间中更加分散，计算式如下：

$$loss_{\cos}=\max[1-\cos(x_i,x_j)] \tag{10}$$

其中，x_i 和 x_j 为 n 段帧经过特征提取后得到的 768 维的特征向量，$i,j\in[1,\cdots,n]$，$1-\cos(x_i-x_j)$为同类余弦相似度。为了加快收敛以及节省算力，对于 n 段帧，在提取其特征后，只取余弦相似度相差最大的两个特征向量计算损失，并进行反向传播，具体过程如下：先选取一个视频的 n 段帧，计算每帧间的余弦相似度差异；取其中的最大值为该视频的损失值，再将批次中所有视频的损失取平均，作为该批次的损失值；在此基础上进行反向传播，从输出层开始，利用链式法则计算出每个神经元权重对损失的贡献，再根据该贡献计算参数权重需要更新的梯度。

本文提出的网络模型采用的损失函数，在注重分类准确率的同时，还可以减小同一视频多个片段之间的帧特征差异。计算公式如下：

$$Loss_{\text{all}}=\lambda loss_{\text{CE}}+\eta loss_{\text{COS}} \tag{11}$$

其中，$Loss_{\text{all}}$表示本文采用的损失函数，$loss_{\text{CE}}$为交叉熵损失函数，λ 为该损失的权值，$loss_{\text{COS}}$为余弦相似度损失函数，η 为该损失的权值($\eta+\lambda=1$)。将交叉损失函数与改进的余弦相似度损失函数结合，以起到联合优化的效果，提高模型特征的表征能力。

3 实验与结果分析

3.1 数据集与实验环境

本文采用的数据集为 UCF101 和 HMDB51。

UCF101 来自 YouTube，是最具挑战性的数据集之一，其中各视频在背景、光照条件、视角、外观和姿态等方面都可能存在较大的差异，共包含 101 个动作类别，每个类别分为 25 组，每组 4—7 个短视频，共 13 320 个视频。本文所用的训练集为其中的 9537 个视频，测试集为其中的 3783 个视频，训练集与测试集的比例约为7∶3。

HMDB51 是一个公共数据集，包含 51 个动作类别，主要为面部动作、一般身体动作、物体交互动作、人类互动动作，共 6766 个视频。本文所用的训练集为其中的 5035 个视频，测试集为其中的 1731 个视频，训练集与测试集的比例约为7∶3。

本实验在 Windows10 64 位操作系统上进行，CPU 为 Intel Xeon Gold 5118 2.30 GHz，使用 GPU 加速运算，GPU 为 NVIDIA Tesla V100，内存为 32 GB，实验代码使用 Python 语言并基于 Pytorch 框架实现。

3.2 实验设置

实验首先使用 OpenCV 提取了视频的所有 RGB 帧，再使用 TV-L1 光流算法计算相邻帧的光流，以便得到实验所需的光流数据。

在训练时，RGB 图像批尺寸为 30，通道数为 3，光流图像批尺寸为 10，通道数为 10，即 x 与 y 方向的连续各 5 帧光流图。为了使模型更具鲁棒性，采用随机梯度下降算法作为优化器，共训练 40 个轮次，初始学习率设置为 0.001；为了抑制过拟合，加速度设置为 0.9，权重衰减率为 0.0005，每 2 轮训练后调整学习率为原来的 0.5 倍。此外，损失函数采用改进的联合损失函数，其中 $loss_{\text{CE}}$的权值 λ 为 0.6，$loss_{\text{COS}}$的权值 η 为 0.4。此外，针对模型过拟合的问题，还采用了以下方法：

1)交叉模态预训练技术，对于本文模型的基础网络，首先实例化图像分类任务中的网络，读取其在 ImageNet 数据集上的预训练权重，再将网络尾部分类层的类别数修改为本文实验所用数据集对应的类别数，并重新初始化分类层参数，从而实现对网络参数的充分寻优。对于 RGB 数据，在修改分类层后可直接加载权重。对于光流数据，通过线性变化将光流数据离散至 0～255 的区间，从而转换至与 RGB 单通道相同的值域；将 ViT 的第一个卷积层的权重进行平均，根据光流数据的通道数复制该平均值；修改 ViT 的第一个卷积层输入通道数，并读取平均后的权重。

2)正则化技术，在 ViT 的层归一化与 Dropout 的基础上，在模型最后一层归一化后添加了一个额外的 Dropout 层，以进一步降低过拟合的影响，Dropout 率设置为 0.5。

3)数据增强技术，包括尺度抖动、角裁剪与随机水平翻转，这些操作有助于提高模型对不同角度不同位置数据的识别能力。例如：尺度抖动会将视频帧的大小随机缩放到一个固定的尺度范围内，以增加模型对不同尺度图像的适应能力；角裁剪在图像的中心或 4 个角落进行裁剪，以避免模型只关注图像的中心；随机水平翻转以 50%的概率随机对图像进行水平翻转。之后再对图像进行归一化和正则化处理，以增加模型对不同角度数据的适应能力。

在测试时，RGB 图像与光流图像的批尺寸都设置为 1，将每个视频所有帧均分为 25 段，从每段中随机抽取一帧共得到 25 帧，将所有图片缩放后再进行中心裁剪，并进行归一化和正则化处理，得到 25 个 224×224 的图像。

为了直观地反映模型的效果，本文主要采用准确率作为算法的评价指标，具体计算式如下：

$$Acc=\frac{(TP+TN)}{(TP+TN+FP+FN)} \tag{12}$$

其中，TP 与 TN 为预测正确的正类与负类，FP 与 FN 为预测错误的正类与负类，准确率为正确预测结果在全部的预测结果中所占的比例。

3.3 采帧方式消融实验

在加载数据集时，需要先对视频进行分段采样，即从帧序列中采帧。为验证采帧方式对模型准确率的影响，本文预设了 4 种采帧方式，包括：1)取视频中间相邻的 3 帧；2)将视频均分为 3 段后，取每段中间的帧；3)将视频均分为 3 段后，对每段都取随机帧；4)将视频均分为 3 段后，对第一段取随机帧，在其余两段中取同样位置的帧。

为了保证实验的可靠性，每种方法的实验均采用相同的模型，Baseline 都为 ViT。为了消除实验的随机性，本文对每种方法均做 3 次实验，取平均值为最终结果。每种方法的结果如表 1 所列，其中将视频均分为 3 段后，取每段中间的帧，即第二种方法效果最好。相比第一种方法，这种方法可以获取更多的帧信息，有助于模型更好地捕捉关键帧和动作细节。相比第三种和第四种方法，该方法选取的帧位置相对固定，有利于消除随机性带来的影响，同时也避免了在视频起始或结束处出现的不规则帧序列。因此，第二种方法在 4 种视频帧采样方案中表现最好，能够提高模型的准确率。

表 1 不同采帧方式在 UCF101 上的 RGB 流准确率对比

Table 1 Comparison of RGB results of different way of selecting frames on UCF101

Way	1	2	3	4
Acc/%	85.43	87.28	86.35	86.12

3.4 决策层融合实验

图 4 给出了 RGB 流与光流用不同权重融合后的准确率曲线，其中横坐标为 RGB 流的权重，纵坐标为融合后的准确率。以 0.1 为间隔改变 4 个网络的 RGB 流权重并进行对比，共进行 40 次实验，可以看到，当 RGB 流的权重为 0.6 时，所有曲线准确率都为最高，因此权重比的改动是有效的。究其原因，主要是由于 RGB 流中的颜色与纹理信息是非常重要的视觉特征，可以很好地表现出不同动作的视觉细节和上下文信息，而光流信息中存在复杂场景或相机运动等干扰，通过充分利用 RGB 流的颜色与纹理信息，能够弥补光流信息中的不足，特别是在加入了 SimAM 注意力机制和联合损失后，准确率都有所提升。

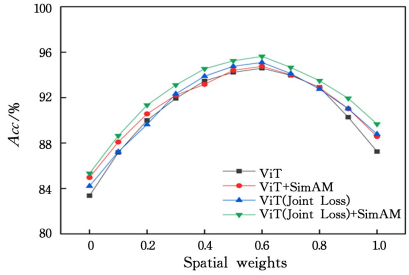

图 4 不同融合权重在 UCF101 数据集上的准确率

Fig. 4 Recognition rate of different fusion weights on UCF101

3.5 不同模型对比实验

行为识别是从连续视频帧中自动识别出人体行为的过程，其本质可以看作是一种图像分类任务。而基于自注意力机制的 ViT 模型，能够捕获像素块之间的长距离关系，从而避免了卷积神经网络中的局部感受野的限制；通过迁移学习的方法，加载图像分类任务的预训练权重，可以减小缺少归纳偏置所带来的影响，因此 ViT 在行为识别任务中有非常出色的表现。

如图 5 所示，在将 Baseline 替换为 ViT 后，模型收敛得更快，收敛后的 Loss 也有明显下降。在更换损失函数与嵌入 SimAM 后，模型的收敛速度进一步提升，并且 Loss 也进一步下降。

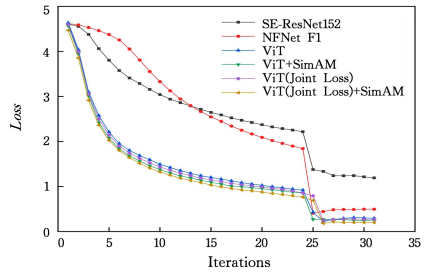

图 5 不同 Baseline 在 UCF101 上的 RGB 流训练 Loss 曲线

Fig. 5 RGB training loss curves for different baselines on UCF101

对于行为识别任务，其重要因素包括行为人的肢体、行为相关物，而无关因素包括动作背景等。模型学习到的重要因素越多，注意到的无关因素越少，模型表现越好。如图 6 所示，在嵌入 SimAM 后，能够有效地增加模型对人的肢体以及行为相关物的注意力增强，减少了动作背景的干扰，从而提高了模型的特征表示能力。

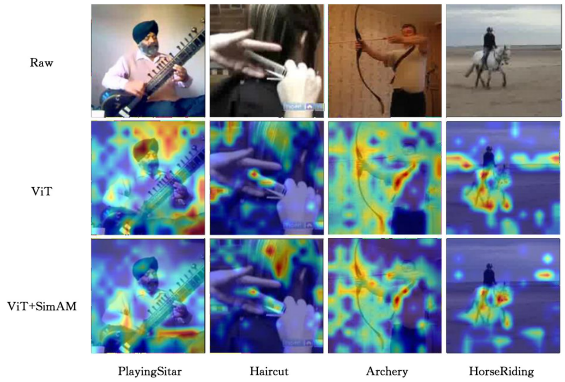

图 6 可视化特征热力图

Fig. 6 Visual feature heat maps

表 2 列出了 3 个深度网络模型在行为识别任务上的表现的比较结果，其中 ViT 的表现最优。SE-ResNet152[14] 与 NF-Net-F1[15] 为基于卷积运算的神经网络，在图像分类任务上表现较好，由于其卷积操作具有平移不变性和局部敏感性等归纳偏置，因此仅需要较少的数据便可以训练出一个模型，但其缺乏对图像全局信息的感知，无法对特征之间的依赖关系进行判断，从而导致对图像的全局信息理解不足。而 ViT 的自注意力机制能够使网络关注到图像中任意块之间的关系，因此可以很好地弥补这一缺点，并充分利用上下文信息对图像的全局依赖关系进行建模。

表 2 不同 Baseline 在 UCF101 上的准确率对比

Table 2 Accuracy comparison of different baselines on UCF101

Baseline	Spatial/%	Temporal/%	Two-Stresm/%	♯param.	GFLOPs
SE-ResNet152	83.06	72.55	89.65	65×10^6	148.06×10^9
NFNet-F1	85.97	83.22	93.19	129.87×10^6	226.8×10^9
ViT	87.28	83.37	94.59	85.7×10^6	228.36×10^9
ViT+SimAM	88.58	84.97	95.35	85.7×10^6	228.36×10^9
ViT(Joint Loss)	88.79	84.21	95.10	85.7×10^6	228.36×10^9
ViT(Joint Loss)+SimAM	**89.67**	**85.33**	**95.63**	$\mathbf{85.7\times10^6}$	228.36×10^9

3.6 时间注意力与分段数对比实验

为了充分提取时序特征，将时间自注意力机制添加到模型中，并对不同的插入位置进行了消融实验。如表 3 所列，当时间注意力机制处于 ViT 的头部时，会过早地融合时序信息与视觉信息，此时模型容易受到时序信息的干扰，导致无关信息的冗余与重要信息的不足，从而使得模型难以理解帧间的关系；当时间注意力机制处于 ViT 的尾部时，可以确保模型在融合时序信息时已经具备足够抽象的视觉语义信息，使得

时间注意力机制能够帮助模型更好地理解输入序列中不同时间帧之间的关系，从而更准确地捕捉到序列中的重要信息，行为识别任务侧重于对帧间关系的理解，因此相比添加在 ViT 的头部或不添加，在 ViT 尾部添加时间自注意力机制能够取得很好的效果。

此外，为了验证分段数对模型准确率的影响，本文分别进行了分段数为 3，6，9 的实验。如表 3 所列，将视频进行分段比不分段时的准确率更高，因此将视频分段能够提高模型对长时序列数据的处理能力，但随着分段数的增加，过多的分段数不仅无法使模型有效地提取时序特征，反而会带来计算量的增加以及训练效率的降低，使得准确率下降。

表 3 不同帧数及是否添加时间注意力在 UCF101 上的准确率对比

Table 3 Comparison of RGB results of different segments and whether to add temporal attention on UCF101

Num of segments	Head/%	End/%	None/%	GFLOPs
1Frame(RGB)	—	—	85.03	17.58×10^9
3Frames(RGB)	88.23	**90.12**	89.67	$52.74(+4.16)\times10^9$
6Frames(RGB)	85.72	88.71	86.55	$105.42(+8.41)\times10^9$
9Frames(RGB)	84.85	87.39	86.36	$158.13(+12.63)\times10^9$
3Frames(Flow)	85.49	**86.27**	85.33	$175.62(+41.69)\times10^9$
3Frames (RGB+Flow)	95.46	**96.12**	95.63	$228.36(+45.85)\times10^9$

3.7 结果对比

本文各阶段的实验结果如表 4 所列，可以看出，在改进的过程中，准确率均有提升，证明了改进的有效性。

表 4 在 UCF101 上的消融验证实验结果

Table 4 Results of ablation validation experiment on UCF101

Model	*Acc*/%
ViT(RGB)	86.35
+Selecting Frames(Way 2)	87.28
ViT(RGB+Flow)	94.23
+Decision Fusion(0.6)	94.59
+SimAM	95.35
+Joint Loss	95.63
+Temporal Attention	96.12

为了验证本文方法的先进性，将本文模型识别结果同其他主流方法进行比较，如表 5 所列，本文模型优于目前的主流算法。

本文方法在 UCF101 和 HMDB51 上相比 TSN 的准确率分别提高了 3.48%和 7.76%，相比 STACNet[16] 分别提高了 1.79%和 4.36%。STACNet 使用 SAM 分配空间注意力，使用 TAM 自适应地区分序列中的关键帧，然后使用 CNN 提取时空特征。相比 TSN 与 STACNet，本文方法首先在网络头部和尾部嵌入了无参注意力模块与时间注意力模块，使用了更高效更全面的注意力机制来评估特征权重，提高了模型对背景干扰与长时序列数据的处理能力；其次，使用了基于预训练 ViT 的时空特征提取模块，从而能够充分捕捉图像块特征之间的空间关系，得到更具有表征性的图像特征向量；然后，优化了训练的损失函数，从而提高了模型的分类性能；最后，采用了决策融合策略，从而充分利用了两种特征所带来的不同优势。因此，本文方法可以有效提高模型在行为识别任务上的精度，具有一定优越性。

相比 TimeSformer，本文方法在 UCF101 和 HMDB51 上的准确率分别提高了 3.69%和 9.92%。TimeSformer 一般输入长时序列视频帧，且在每层 Transformer 的空间特征提取后都进行了时间特征的提取，使其需要大量的计算资源来进行训练和推理。此外，由于 TimeSformer 主要侧重于对单个视频帧的时空特征建模，导致其无法捕捉到长时序列帧间关系的复杂信息。虽然独特的注意力机制使其在 Kinetics-400 等大数据集上有较好的表现，但由于每层都进行时间自注意力的运算，导致该模型较深较大、泛化性较差，在 UCF101 及 HMDB51 上的表现不够好。本文方法对帧间关系的建模与特征的提取更为全面，在小数据集上的较优结果表明本文算法的泛化性较好，能够在一些资源较为受限的场景下进行应用。

表 5 不同算法在 UCF101 和 HMDB51 上的准确率对比

Table 5 Comparison of accuracy of different algorithms on UCF101 and HMDB51

(%)

Model	UCF101	HMDB51
TSN	92.64	65.74
3D ResNeXt-101[17]	91.20	—
VideoMAE[18]	91.30	62.60
TimeSformer[12]	92.43	63.58
CSCNN[19]	92.70	64.50
HAR-depth[20]	93.00	69.70
STACNet[16]	94.33	69.14
TVBN-ResNeXt[21]	94.60	70.40
HAM-CNN[22]	94.70	70.90
spatio-temporal STFT[23]	94.70	71.50
LIGAR[24]	94.85	—
BifurcatedNet[25]	94.90	72.10
STIAM[26]	94.90	—
TS-CNN[27]	94.90	70.80
DAM[28]	95.70	71.80
ViTSN	96.12	73.50

结束语 本文提出了一种基于改进双流视觉 Transformer 的行为识别模型。该模型利用 SimAM 降低了动作背景的干扰；利用 ViT 与时间注意力机制提取了图像空间特征与帧间的时序特征；利用改进的损失函数增强了模型的特征提取能力；利用决策层融合了光流模态与 RGB 模态数据特征，从而充分提取了视频的时空特征。在 UCF101 与 HMDB51 数据集上与现有方法进行比较，结果表明本文模型的识别准确率更好，具有一定的优越性。

后续的研究工作包括如何进一步对不同细粒度的行为进行识别，以及如何将本文模型应用于时空行为检测任务中，以获得更多的应用结果。

参 考 文 献

[1] MA Y X, TAN L, DONG X, et al. Action Recognition for Intelligent Monitoring[J]. Journal of Image and Graphics, 2019, 24(2): 282-290.

[2] CHU J H, ZHANG S, LV W. Driving Behavior Analysis Algorithm Based on Convolutional Neural Network[J]. Laser & Optoelectronics Progress, 2020, 57(14): 141018.

[3] SUN Q, JI G L, ZHANG J. Non-Local Attention Based Genera-

tive Adversarial Network for Video Abnormal Event Detection [J]. Computer Science, 2022, 49(8): 172-177.
[4] MIAO Q G, XIN W T, LIU R Y, et al. Graph Convolutional Skeleton-Based Action Recognition Method for Intelligent Behavior Analysis[J]. Computer Science, 2022, 49(2): 156-161.
[5] KARPATHY A, TODERICI G, SHETTY S, et al. Large-Scale Video Classification with Convolutional Neural Networks[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington DC: IEEE Computer Society, 2014: 1725-1732.
[6] HE M, ZHU C, HUANG Q, et al. A Review of Monocular Visual Odometry[J]. The Visual Computer, 2020, 36(5): 1053-1065.
[7] SIMONYAN K, ZISSERMAN A. Two-stream Convolutional Networks for Action Recognition in Videos[C]// The 27th International Conference on Neural Information Processing Systems. Montreal: NIPS'14, 2014: 568-576.
[8] NEIMARK D, BAR O, ZOHAR M, et al. Video Transformer Network[C]// Proceedings of IEEE/CVF International Conference on Computer Vision Workshops[C]// Piscataway, NJ: IEEE Press, 2021: 3156-3165.
[9] ARNAB A, DEHGHANI M, HEIGOLD G, et al. ViViT: A Video Vision Transformer[C]// Proceedings of IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2021: 6816-6826.
[10] FAN H, XIONG B, MANGALAM K, et al. Multiscale Vision Transformers[C]// Proceedings of IEEE/CVF International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2021: 6804-6815.
[11] YAN S, XIONG X, ARNAB A, et al. Multiview Transformers for Video Recognition[C]// 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE Press, 2022: 3333-3343.
[12] BERTASIUS G, WANG H, TORRESANI L. Is Space-time Attention All You Need for Video Understanding? [C]// 2021 ICML. Virtual: IEEE Press, 2021.
[13] YANG L, ZHANG R Y, LI L, et al. Simam: A Simple, Parameter-free Attention Module for Convolutional Neural Networks [C]// Proceedings of the 38th International Conference on Machine Learning. New York: PMLR, 2021: 11863-11874.
[14] HU J, SHEN L, SUN G. Squeeze-and-excitation Networks [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City: IEEE, 2018: 7132-7141.
[15] BROCK A, DE S, SMITH S L, et al. High-Performance Large-Scale Image Recognition without Normalization[C]// Proceedings of the 38th International Conference on Machine Learning. New York: PMLR, 2021: 1059-1071.
[16] LIU S, MA X, WU H, et al. An End to End Framework with Adaptive Spatio-Temporal Attention Module for Human Action Recognition[J]. IEEE Access, 2020, 8: 47220-47231.
[17] SHALMANI S M, CHIANG F, ZHENG R. Efficient Action Recognition Using Confidence Distillation[C]// The 26th International Conference on Pattern Recognition. Montreal: IEEE, 2022: 3362-3369.
[18] TONG Z, SONG Y, WANG J, et al. Videomae: Masked Autoencoders Are Data-Efficient Learners for Self-Supervised Video Pre-Training[J]. arXiv: 2203.12602, 2022.
[19] YI Z W, SUN Z H, FENG J C, et al. Channel Separable Convolutional Neural Network for Action Recognition[J]. Journal of Signal Processing, 2020, 36(9): 1497-1502.
[20] SAHOO S P, ARI S, MAHAPATRA K, et al. HAR-Depth: A Novel Framework for Human Action Recognition Using Sequential Learning and Depth Estimated History Images[J]. IEEE Transactions on Emerging Topics in Computational Intelligence, 2020, 5(5): 813-825.
[21] HU Z P, ZHANG R X, ZHANG X, et al. TVBN-ResNeXt: End-to-End Fusion of Space-Time Two-Stream Convolution Network for Video Classification[J]. Journal of Signal Processing, 2020, 36(1): 58-66.
[22] WANG Z Q, ZHANG W Q, ZHANG L, et al. Human Behavior Recognition with High-Order Attention Mechanism[J]. Journal of Signal Processing, 2020, 36(8): 1272-1279.
[23] KUMAWAT S, VERMA M, NAKASHIMA Y, et al. Depthwise Spatio-Temporal STFT Convolutional Neural Networks for Human Action Recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(9): 4839-4851.
[24] IZUTOV E. LIGAR: Lightweight General-Purpose Action Recognition[J]. arXiv: 2108.13153, 2021.
[25] ZHANG J, HU H, LIU Z. Appearance-and-Dynamic Learning with Bifurcated Convolution Neural Network for Action Recognition[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 31(4): 1593-1606.
[26] PAN N, JIANG M, KONG J. Human Action Recognition Algorithm Based on Spatio-Temporal Interactive Attention Model [J]. Laser& Optoelectronics Progress, 2020, 57(18): 181506.
[27] ZHANG W Q, WANG Z Q, ZHANG L. Human Action Recognition Combining Sequential Dynamic Images and Two-Stream Convolutional Network[J]. Laser & Optoelectronics Progress, 2021, 58(2): 0210007.
[28] LI C, HE M, DONG C, et al. Action Recognition Model of Directed Attention Based on Cosine Similarity[J]. Journal of System Simulation, 2024, 36(1): 67-82.

LEI Yongsheng, born in 1999, postgraduate. His main research interests include digital forensics and so on.

DING Meng, born in 1980, master, associate professor, postgraduate supervisor. His main research interests include digital forensics and video processing.

（责任编辑：喻藜）