DOI:10.13203/j.whugis20190101 文章编号:1671-8860(2021)08-1241-06

引导式的卷积神经网络视频行人动作分类改进方法

毛 琳[1] 陈思宇[1] 杨大伟[1]

1 大连民族大学机电工程学院,辽宁 大连,116600

摘 要:如何提升网络模型对时域信息的理解能力,是基于3D卷积神经网络视频行人动作分类方法需要解决的问题之一。提出一种主导层优化模块,在网络训练过程中,利用当前时域动态信息学习能力最强的卷积层作为主导层来引导网络权重参数的更新,使各卷积层对动态信息的学习能力逐渐增强,从而改进卷积神经网络模型对时域动态信息的理解能力。仿真结果显示,添加主导层优化模块后的ResNeXt-50网络与ResNeXt-101网络在UCF-101和HMDB-51数据库上的训练收敛速度都有所增加,测试结果的准确率均有不同程度提升。

关键词:视频行人动作分类;动态信息学习能力;引导优化;3D卷积神经网络;时域动态信息理解能力

中图分类号:P237 **文献标志码**:A

视频行人动作分类属于视频理解领域,是一种获取连续帧图像信息并进行动作分类的技术。相较于单帧图像的场景、目标、轮廓等静态信息而言,连续帧图像所包含的时域动态信息的复杂程度更高,理解与学习难度更大。因此,视频理解领域相关技术有待进一步提高。

基于神经网络的视频行人动作分类主要有两种方法:一种是利用3D或2D+1D卷积核构成结构各异的卷积神经网络[1-4],直接对视频数据进行动作信息的提取与学习;另一种双流网络需要在同一个神经网络下对图像原始信息与图像光流信息进行不同分支的训练,模型的分类决策由两个分支共同影响[5],该方法的准确率较高,但是在网络训练之前需要花费大量时间来预处理图像的光流信息。然而,部分学者通过实验发现这两种主流方法并没有使神经网络充分地学习到视频数据的动作信息。双流网络的高准确率是由于光流信息能够很好地保证图像外观的不变性,并不能说明它可以表达运动信息[6]。卷积神经网络因为卷积运算的特性,在模型的训练过程中,更偏向于将视频中的静态信息作为学习重心,对动态信息的学习还有待加强[7-8]。

为了改进神经网络模型在视频行人动作分类任务中的性能,许多学者从不同的切入点着手进行研究。Luo等[9]提出一种图蒸馏法来提炼多个有效模型的参数,结合弱监督与局部多模态训练方式来辅助模型的学习,以此提升网络性能。但方法架构的设计还处于初级阶段,需要结合更高级的跨领域适应算法来提升设计有效性。Wang等[10]用非局部运算来捕获连续图像帧之间的局部依赖性,通过非局部神经网络的结构特性来补充数据在逐层传递过程中损失的依赖信息,提升模型对视频数据中各种动作行为局部时域相关性的学习能力。该方法的灵活性较强,但面对特定任务时还需要进行相应预处理,才能使网络模型达到预期效果。Diba等[11]提出深度时序线性编码网络,对视频数据中不同位置的特征进行融合编码,用改善特征表示的方法来提升网络对时间特征的理解能力,但网络的特征编码仅针对时间信息,在如何聚合时间与空间信息上还可以改良。

为了提升卷积神经网络(convolutional neural networks,CNNs)在视频行人动作分类任务中的性能,本文针对CNNs模型动态信息理解能力不足的问题设计了一种主导层优化模块,利用主导层引导各个卷积层权重参数的更新方向,根据网

收稿日期:2019-05-13

项目资助:辽宁省自然科学基金(20170540192,20180550866)。

第一作者:毛琳,博士,副教授,主要从事多传感器信息融合、目标跟踪研究。maolin@dlnu.edu.cn

络所学特征与原始数据的动态信息差异进行偏移，优化网络对时域动态信息的学习与理解能力，提升模型的分类准确度。经仿真实验证明，在结合3D残差卷积神经网络[12-13]进行仿真训练的情况下，引导式卷积神经网络优化模块对UCF-101[14]与HMDB-51[15]数据集测试结果的Top-1和Top-5准确率（Top代表网络模型输出分类结果时的概率高低序列，Top-1准确率指网络模型输出结果中的最高序列为正确结果，Top-5准确率代表输出结果的前5序列中包含的正确结果）分别拥有不同程度的提升。

1 主导层优化模块

该模块将卷积层数据与输入层数据的动态信息进行相似度对比，对各卷积层的动态信息学习能力进行评估，动态信息学习能力最强的卷积层被确立为主导层。计算主导层与输入层之间时域动态信息的量化差异，将该差异值作为约束条件添加到损失函数，反馈优化各卷积层权重参数的梯度，从而影响卷积层权重参数的优化更新方向。主导层优化模块的逻辑框图如图1所示。

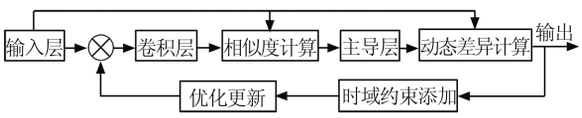


图1 主导层优化模块逻辑框图

Fig.1 Dominant Layer Optimization Module Logic Block Diagram

卷积层权重参数决定了CNNs模型对输入数据内场景、目标、轮廓、动作等各种信息的学习与理解能力，参数的更新梯度在训练中的微小变化，会使模型对各种信息的理解程度有所改变。主导层优化模块的任务是利用主导层的动态信息学习能力，引导并优化各卷积层权重参数的梯度更新，提升卷积层对动态信息的学习能力，改进网络模型对时域动态信息的理解能力。因此，主导层的有效性与时域约束的准确性是对权重参数进行优化更新的重要保障。

在主导层优化模块的设计预期中，各层数据动态信息在训练时产生的变化如图2所示。

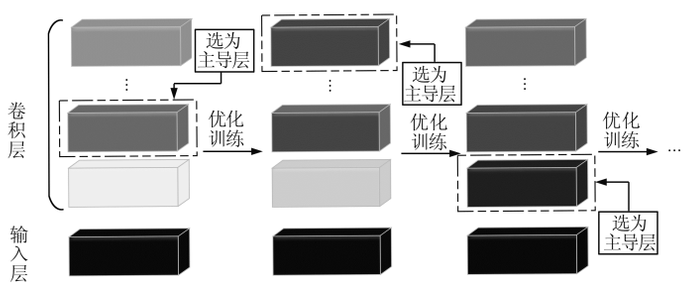


图2 模块对卷积层时域动态信息的影响及选择主导层的概念图

Fig.2 Conceptual Diagram About the Influence of the Module on Temporal Information of Convolutional Layer and the Selection of Dominant Layer

图2中，方块代表CNNs输入层与若干中间层数据的动态信息，方块颜色深浅的相近程度代表了层间动态信息的相似程度，动态信息与输入层最相似的卷积层将被确立为主导层。模块通过在整体上缩小卷积层与输入层之间时域动态信息差异的方式来达到优化训练的效果，提升网络模型对时域信息的理解能力。

1.1 主导层确立

由于链式法则的传递性，权重参数的更新对网络模型各方面（对不同种类信息的学习能力）性能产生的影响会因信息的逐层传递而变得难以把握。为了确保模块的优化能改进网络的动态信息学习能力，在每次训练周期内，模块仅从卷积层中确立一个最有代表性（动态信息学习能力最强）的主导层作为时域约束的计算参照。通过模块的引导式架构，使主导层的动态信息学习能力对所有卷积层权重参数的更新梯度产生一定的偏向性影响，让网络模型在保障性能的前提下适当增强对时域动态信息的注意力[16-17]，避免模型由于引导式优化可能产生的过拟合现象。

通过对输入层数据与各卷积层数据的时域动态信息进行相似度计算，将动态信息相似程度的高低等价于卷积层对输入数据动态信息学习能力的强弱，以此筛选出学习能力最强的卷积层作为优化模块的主导层。

用卷积层特征数据动态信息集合T^C与输入层原始数据动态信息集合T的数值化差异d来等价数据组相似程度的高低，确立差异结果d最小，即与输入层原始数据动态信息相似度最高的卷积层为主导层。差异计算过程为：

$$d=\left|\delta(T)-\delta\left(T^{C}\right)\right| \tag{1}$$

式中,δ代表对动态信息集合的数值化表示。为了获取视频数据中更宽泛的全局动态信息,防止有效信息被剔除,由帧间差值计算得出各层数据的动态信息集合。对两组集合数值化表示的具体计算为:

$$\begin{cases}\delta\left(T^{C}\right)=\det\left(T^{C}+E\right)\\ \delta(T)=\det(T+E)\end{cases} \tag{2}$$

通过这种动态信息提取方式,避免在模块引导优化的过程中,网络对局部信息注意力的过度提高导致模型整体性能下降,确保模块的实用性。

卷积层特征数据动态信息集合T^{C}的具体计算过程可表示为:

$$\begin{cases}T^{C}=\left\{t_1^{C},t_2^{C},t_3^{C}\cdots t_{n-1}^{C}\right\}\\ t_i^{C}=u_{i+1}^{C}-u_i^{C},1\leqslant i\leqslant n-1\end{cases} \tag{3}$$

式中,u^{C}代表特征数据;n代表数据组的总帧数;式(3)中第2式表示动作信息的计算过程;i用来代表数据组的任意某一帧。输入层原始数据动作信息集合T的计算过程与T^{C}相同。

对于各个卷积层内数据通道数不相同的情况,本文对通道数量更大的卷积特征数据按整体比例随机取样,以取样组的期望值作为数值化计算的输入数据,确保各个数据组分布的一致性。

1.2　权重参数优化更新

主导层优化模块需要量化主导层动态信息集合$T^{C'}$与T这两组动态信息集合之间的距离,将量化结果作为时域约束添加到网络的损失函数中。时域约束与模型准确度成本结合后的损失函数,使两者共同对网络的梯度造成影响,导致卷积层权重参数W的更新在保证模型准确度提升的同时,确保$T^{C'}$与T的距离被拉近,减小主导层所学特征与输入层原始数据的时域差异。

添加时域约束后的损失函数及权重参数W更新运算的具体表示如下:

$$\begin{cases}L=L_p\left(l,\hat{l}\right)+\lambda D\left(T,T^{C'}\right)\\ W=\arg\min\limits_{T^{C'}}\left\{L_p\left(l,\hat{l}\right)+\lambda D\left(T,T^{C'}\right)\right\}\end{cases} \tag{4}$$

式中,$L_p\left(l,\hat{l}\right)$为输入样本真实值与网络模型分类输出值构成的损失函数;λ为约束权重;$D\left(T,T^{C'}\right)$代表时域约束,即输入层动态信息集合T与主导层动态信息集合$T^{C'}$之间的量化距离。

由于原始数据与特征数据并不属于同一数据空间,常见的差异计算方法会忽视两组数据之间的跨空间性,导致计算结果不够可靠。本文借鉴迁移学习中多领域自适应的相关技巧,使用最大均值差异(maximum mean discrepancy,MMD)算法[18-20]对两组跨空间动态信息的距离进行稳健的量化估计。

MMD算法能够将两组数据利用某一样本空间的连续函数集F进行映射,根据映射结果的差异大小来判断它们是否属于同一分布。如何选择F是计算结果准确与否的关键,两组样本在函数集中得到的最大映射差异为MMD,如果MMD足够小,则代表两组样本属于同一分布。MMD算法及其经验估计被定义为:

$$\begin{cases}\mathrm{MMD}[F,p,q]=\sup\limits_{f\in F}\left(E_p[f(x)]-E_q[f(y)]\right)\\ \mathrm{MMD}[F,X,Y]=\sup\limits_{f\in F}\left(\dfrac{1}{m}\sum\limits_{i=1}^{m}f(x_i)-\dfrac{1}{n}\sum\limits_{i=1}^{n}f(y_i)\right)\end{cases} \tag{5}$$

式中,p与q被假设为波莱尔概率分布;X与Y是分布p与q经过独立同分布采样得到的大小分别为m与n的数据集;f是函数集F中的某个函数。

主导层优化模块利用再生核希尔伯特空间的MMD算法,将两组不同空间的动作信息集合利用核函数统一映射到希尔伯特空间,以映射后的经验估计作为两组数据的距离度量结果,将其等同于两组跨空间时域动作信息的量化差异。该映射空间的完备性与正则性能够确保计算结果的稳健程度,并且在样本数据量较大时计算出的数据差异不会有过大的偏差。

再生核希尔伯特空间MMD算法的函数集F中每一个映射函数f都通过核函数φ将输入数据映射到再生核希尔伯特空间,其公式表达为:

$$\begin{cases}f(\boldsymbol{x})=\left\langle f\varphi(\boldsymbol{x})\right\rangle_H\\ \varphi(\boldsymbol{x})=\exp\left[-\|\boldsymbol{x}-\boldsymbol{x}'\|^2/\left(2\sigma^2\right)\right]\end{cases} \tag{6}$$

式中,φ代表数据用高斯核进行映射;H代表希尔伯特空间;$\boldsymbol{x}'$代表输入数据的转置;σ为样本标准差。

以输入层原始数据的动态信息集合T与主导层特征数据的动态信息集合$T^{C'}$作为输入MMD算法的跨空间数据,将两组时域动态信息集合内各样本最大均值差异的期望值作为时域约束,以此来完成模块的反馈闭环传递。时域约束的计算为:

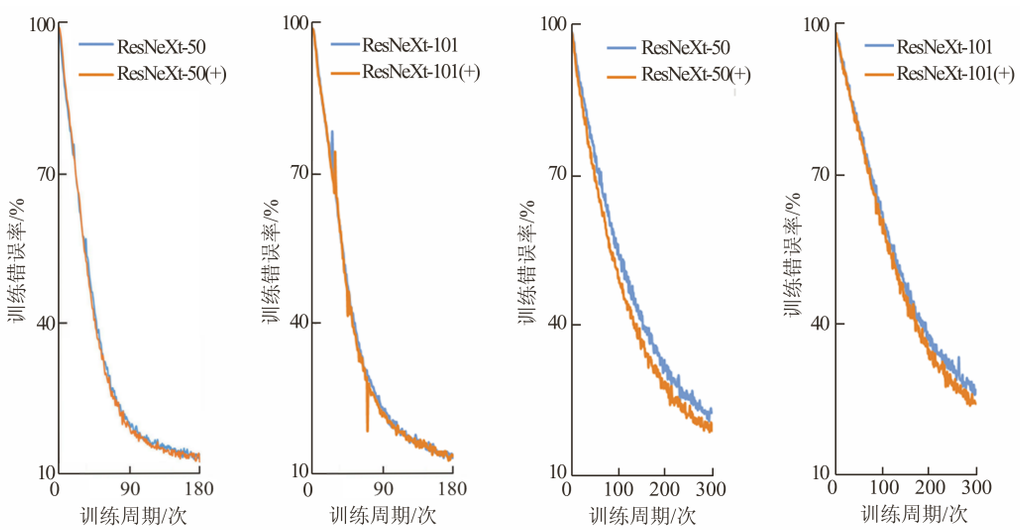
$$D\left(T,T^{C'}\right)=\frac{1}{n-1}\left\|\sum_{i=1}^{n-1}\varphi\left(t_i\right)-\sum_{i=1}^{n-1}\varphi\left(t_i^{C'}\right)\right\|_H^2 \tag{7}$$

2 仿真结果与分析

本文的仿真实验选择在视频行人动作分类中较为常用的数据集UCF-101与HMDB-51，合计有152个动作类和近20 000个动作实例，样本数量充足且训练样本占整个数据集的70%，另外30%的数据为测试样本，该分布能确保模型对数据的有效学习与验证。

本文选择结合3D卷积核的残差网络作为仿真的基底网络。由于主导层的确立需要对各卷积层进行筛选，当所用基底网络的卷积层数量较多时，虽然筛选的计算过程繁重，但能够使优化模块对主导层的选择更稳健。同时，深层网络在传递梯度时的计算复杂度更高，可以避免模块的优化仅针对某个卷积层产生效果，而无法影响其他层的权重参数，降低模型在训练时过早陷入局部最优的概率。为了增加模型的实用性，本文选用具备卷积分组计算的ResNeXt网络[21]，它能够在保证模型性能的同时减少一定的参数量。由于网络的训练成本将随着层数的加深而大幅度提升，网络性能的提升却较有限，因此，本文仅使用50和101层的ResNeXt网络进行仿真实验。

仿真实验在有8张英伟达1080Ti显卡的图形工作站上运行程序，由于工作站内存有限，在确立主导层时，本文预先从网络中随机选出部分卷积层，再对其进行筛选与确立。为了防止模型参数的更新受到其他信息的干扰，实验时并未使用任何预训练模型。两个网络对不同数据库进行训练时的训练错误率曲线如图3所示。



(a) UCF-101 (b)HMDB-51

图3 不同数据集训练过程中的训练错误率曲线图

Fig.3 Training Error Rate Graphs During Different Database Training

图3中，带有“+”号的表示该模型在原始网络的基础上添加了主导层优化模块。由图3可知，两个网络在添加了主导层优化模块的情况下，对不同数据库训练所得的错误率都有一定程度的下降，表明模块加速了网络训练的收敛过程。虽然模块对网络训练过程有影响，但训练错误率曲线的整体趋势并没有较大的改变，可以看出，模块对网络的引导优化没有使模型的性能产生较大偏差。由于网络层数不同导致网络权重参数量的不同，因此，ResNeXt-50网络的训练错误率变化幅度大于ResNeXt-101网络的。

UCF-101数据集上的网络模型的测试结果如表1所示，3组数据中取均值后得出结果。从表1可见，添加主导层优化模块后，ResNeXt-50的Top-1准确率提升了0.9%，Top-5准确率提升了1.5%；ResNeXt-101的Top-1准确率提升了0.8%，Top-5准确率提升了1.2%。由此可知，本模块能够给网络的性能带来一定程度的优化。

表1 在不同数据集测试集上的正确率/%

Tab.1 Accuracies on Different Data Sets/%

模型	UCF-101		HMDB-51	
	Top-1	Top-5	Top-1	Top-5
ResNeXt-50	47.0	69.2	18.5	43.2
ResNeXt-50(+)	47.9	70.7	20.0	43.6
ResNeXt-101	45.8	68.6	18.6	43.3
ResNeXt-101(+)	46.6	69.8	19.9	43.5

不同网络模型在HMDB-51数据集上的测试结果见表1，从3组数据中取均值后得出结果。优

化后,ResNeXt-50的Top-1准确率提升了1.5%,Top-5准确率提升了0.4%;ResNeXt-101的Top-1准确率提升了1.3%,Top-5准确率提升了0.2%。可见,两种准确率各有不同程度的提升。由此可以了解到,在用HMDB-51数据集进行实验时,主导层优化模块显著提升了模型的精确度,但对网络泛化能力的提升并不明显,还需要通过其他方法来完善模块对泛化能力的优化。

3　结　语

为了提升CNNs模型在视频行人动作分类任务中的性能,本文提出一种主导层优化模块。在训练过程中,模块的引导式优化利用权重参数的更新逐渐增强各卷积层的动态信息学习能力,从而改进网络模型对时域动态信息的理解能力。仿真实验结果显示,本模块优化过的ResNeXt-50与ResNeXt-101提升了不同视频行人动作分类数据库的检测准确率,但模块对网络泛化能力的提升还有待加强。

在以后的工作中,将结合卷积神经网络的结构细节与高阶差异估计算法对动态信息的广义化表达进行更深入的研究,进一步提升网络模型的性能。

参　考　文　献

[1] Tran D, Bourdev L, Fergus R, et al. Learning Spatiotemporal Features with 3D Convolutional Networks[C]// 2015 IEEE International Conference on Computer Vision, Santiago, Chile, 2015

[2] Tran D, Wang H, Torresani L, et al. A Closer Look at Spatiotemporal Convolutions for Action Recognition [C]// IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018

[3] Pei Songwen, Yang Baoguo, Gu Chunhua. Research on Video Stream Classification Using 3D ConvNet Ensemble Fusion Model [J]. *Journal of Chinese Computer Systems*, 2018, 39(10):2 266-2 270(裴颂文, 杨保国, 顾春华. 融合的三维卷积神经网络的视频流分类研究[J]. 小型微型计算机系统, 2018, 39(10): 2 266-2 270)

[4] Wu Peiliang, Yang Xiao, Mao Bingyi, et al. A Perspective-Independent Method for Behavior Recognition in Depth Video via Temporal-Spatial Correlating [J]. *Journal of Electronics and Information Technology*, 2019, 41(4): 904-910 (吴培良, 杨霄, 毛秉毅, 等. 一种视角无关的时空关联深度视频行为识别方法[J]. 电子与信息学报, 2019, 41(4): 904-910)

[5] Simonyan K, Zisserman A. Two-Stream Convolutional Networks for Action Recognition in Videos [C]// Advances in Neural Information Processing Systems, Montreal, Canada, 2014

[6] Sevilla-Lara L, Liao Y, Güney F, et al. On the Integration of Optical Flow and Action Recognition [C]// German Conference on Pattern Recognition, Springer, Cham, 2018

[7] Huang D A, Ramanathan V, Mahajan D, et al. What Makes a Video a Video: Analyzing Temporal Information in Video Understanding Models and Datasets[C]// IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018

[8] Xiong Hanjiang, Zheng Xianwei, Ding Youli, et al. Semantic Segmentation of Indoor 3D Point Cloud Model Based on 2D-3D Semantic Transfer[J]. *Geomatics and Information Science of Wuhan University*, 2018, 43(12): 2 303-2 309 (熊汉江, 郑先伟, 丁友丽, 等. 基于2D-3D语义传递的室内三维点云模型语义分割[J]. 武汉大学学报·信息科学版, 2018, 43(12): 2 303-2 309)

[9] Luo Z, Hsieh J T, Jiang L, et al. Graph Distillation for Action Detection with Privileged Modalities [C]// European Conference on Computer Vision, Munich, Germany, 2018

[10] Wang X, Girshick R, Gupta A, et al. Non-local Neural Networks[C]// IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018

[11] Diba A, Sharma V, Van Gool L. Deep Temporal Linear Encoding Networks[C]// IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017

[12] He K, Zhang X, Ren S, et al. Identity Mappings in Deep Residual Networks[C]//European Conference on Computer Vision, Amsterdam, Netherlands, 2016

[13] Hara K, Kataoka H, Satoh Y. Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and Imagenet?[C]// IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018

[14] Khurram S, Amir Z, Mubarak S. UCF-101: A Dataset of 101 Human Action Classes from Videos in the Wild [EB/OL]. (2012-12-01) [2019-05-13]. https://www. crcv. ucf. edu/papers/UCF101_CRCV-TR-12-01.pdf

[15] Kuehne H, Jhuang H, Garrote E, et al. HMDB: A Large Video Database for Human Motion Recognition[C]// International Conference on Computer Vision, Barcelona, Spain, 2011

[16] Li Rui, Shen Yuqi, Jiang Jie, et al. Temporal and Spatial Characteristics of Hotspots in Public Map Service[J]. *Geomatics and Information Science of Wuhan University*, 2018, 43(9): 1 408-1 415 (李锐, 沈雨奇, 蒋捷, 等. 公共地图服务中访问热点区域的时空规律挖掘[J]. 武汉大学学报·信息科学版, 2018, 43(9): 1 408-1 415)

[17] Hu Tao, Zhu Xinyan, Guo Wei, et al. A Moving Object Detection Method Combining Color and Depth Data[J]. *Geomatics and Information Science of Wuhan University*, 2019, 44(2): 276-282 (胡涛, 朱欣焰, 吴维, 等. 融合颜色和深度信息的运动目标提取方法[J]. 武汉大学学报·信息科学版, 2019, 44(2): 276-282)

[18] Borgwardt K M, Gretton A, Rasch M J, et al. Integrating Structured Biological Data by Kernel Maximum Mean Discrepancy[J]. *Bioinformatics*, 2006, 22(14): e49-e57

[19] Long M, Cao Y, Wang J, et al. Learning Transferable Features with Deep Adaptation Networks[C]// The 32nd International Conference on Machine Learning, Lille, France, 2015

[20] Long M, Zhu H, Wang J, et al. Deep Transfer Learning with Joint Adaptation Networks[C]// The 34th International Conference on Machine Learning, Sydney, Australia, 2017

[21] Xie S, Girshick R, Dollár P, et al. Aggregated Residual Transformations for Deep Neural Networks [C]// IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017

A Guided Method for Improving the Video Human Action Classification in Convolutional Neural Networks

MAO Lin[1] *CHEN Siyu*[1] *YANG Dawei*[1]

1 College of Mechanical and Electronic Engineering, Dalian Minzu University, Dalian 116600, China

Abstract: Objectives: In order to improve the ability of convolutional neural networks (CNNs) of understanding temporal dynamic information, this paper proposes a dominant layer optimization module. **Methods:** The new module uses the dominant layer to guide and optimize the update gradient of convolutional layer weights, and assist the difference estimation with the maximum mean difference algorithm of a reproducing Hilbert space. **Results:** In continuous training, the network can improve the learning ability of temporal dynamic information, and the dynamic information similarity between the features learned by convolutional layer and the input data is also increased. **Conclusions:** This module enhances the performance of the CNNs model on video human action classification and achieves improvements to the network.

Key words: video human action classification; dynamic information learning ability; guided optimization; 3D convolutional neural networks; temporal domain dynamic information understanding ability

First author: MAO Lin, PhD, associate professor, specializes in the multi-sensor information fusion and target tracking. E-mail: maolin@dlnu.edu.cn

Foundation support: The Natural Science Foundation of Liaoning Province(20170540192, 20180550866).

引文格式: MAO Lin, CHEN Siyu, YANG Dawei.A Guided Method for Improving the Video Human Action Classification in Convolutional Neural Networks[J].Geomatics and Information Science of Wuhan University, 2021, 46(8):1241-1246.DOI:10.13203/j.whugis20190101(毛琳, 陈思宇, 杨大伟. 引导式的卷积神经网络视频行人动作分类改进方法[J]. 武汉大学学报·信息科学版, 2021, 46(8):1241-1246.DOI:10.13203/j.whugis20190101)